EPFL

Thesis n° 11248

From Pixels to Attention: A Computational Perspective On Gaze Understanding in the Wild

Presented on 30th March 2026

School of Engineering
L'IDIAP Laboratory
Doctoral program in Electrical Engineering

for the award of the degree of Docteur ès Sciences (PhD)

by

Samy TAFASCA

Accepted on the jury's recommendation

Prof. A. Billard, jury president
Dr J.-M. Odobez, thesis director
Dr F. Brémond, examiner
Prof. A. A. Salah, examiner
Dr M. Salzmann , examiner

École
polytechnique
fédérale
de Lausanne

2026

To my father, without whom this thesis would not exist.

Acknowledgements

I would first like to express my deepest gratitude to my advisor, Jean-Marc, for his guidance, trust, and intellectual support throughout this PhD journey. His vision, rigor, and willingness to grant me scientific independence were instrumental in shaping both this work and my development as a researcher.

I am grateful to the members of my thesis jury — Prof. Aude Billard, Dr. François Brémond, Prof. Albert Ali Salah, and Dr. Mathieu Salzmann — for the time they devoted to reading this manuscript, for their insightful feedback, and for the stimulating discussions that helped strengthen this work.

I would also like to thank my lab members and the broader institutional environments at Idiap and EPFL for providing a supportive and intellectually vibrant setting. In particular, I am grateful to my teammates Anshul, Rémy, Pierre, Michael, Arya, Mitch, Mattia, and Marco for fruitful discussions, collaboration, and shared moments throughout these years. I am equally appreciative of the many friendships formed along the way, including — but not limited to — Victor, Eklavya, Mingchi, Andrei, Fabio, Utsav, Zineb, Driss, Amir, Karl, Roberto, Tilak, Neha, and Darya. Finally, I wish to acknowledge the invaluable support of the Idiap IT helpdesk and technical staff, whose responsiveness and expertise ensured that research could continue smoothly, even when systems failed or last-minute constraints arose.

I would also like to acknowledge my external research partners involved in the broader interdisciplinary projects to which this work contributed, including Marie, Nada, Thomas, Nicolo, Thibaut, and Miranda. These collaborations brought perspectives beyond my immediate research environment and underscored the importance of cross-disciplinary dialogue in addressing complex scientific questions.

I am thankful to my industry collaborators and colleagues met during my internship at NVIDIA, in particular Shalini, Michael, Wookie, Christian, Tianye, Amrita, Jonghyun, Omer, and Koki, for exposing me to different perspectives on research, engineering, and real-world impact. I am also grateful for the friendships formed during this time, including Shankar, Yihang, Sanjev, and Susav.

On a more personal note, I wish to thank my parents for their constant presence and support

throughout this journey. I am equally grateful to my sister Salma, who, at this point, has likely absorbed enough about gaze and attention to write a paper of her own. Last but not least, I am grateful to Mahmoud and Sepehr for their early support, and for remaining a constant source of encouragement throughout the years.

Finally, I would like to acknowledge myself — for persisting when things were difficult, for finding paths forward when doors were closed, for continuing to believe in this work when certainty was scarce, and for not giving up even when the number of supporters was running dry. This thesis is as much a scientific contribution as it is a testament to perseverance.

Lausanne, December 31, 2025 Samy Tafasca

Abstract

Gaze is one of the most revealing and least understood windows into human cognition. Where people look—toward an object or another person—encapsulates attention, coordination, and interaction. In computer vision, however, gaze has long been reduced to a geometric regression problem: predicting a point on an image plane. This formulation captures spatial direction but overlooks the broader utility of gaze as a cue for understanding behavior. This work aims to make gaze understanding more informative and actionable by improving accuracy, scalability, and interpretability through richer geometric, semantic, and uncertainty modeling.

This thesis redefines gaze following as a bridge between vision and social understanding. It develops computational models that reason not only about gaze target locations, but about the attentional relationships linking people, objects, and their environment. Across data, geometry, semantics, and uncertainty, it presents a unified view of gaze as a structured reasoning problem situated in complex, natural scenes.

At the data level, we introduce new benchmarks that capture multiple layers of visual and social context in natural settings. At the modeling level, we propose several architectural advances that (i) couple spatial geometry with semantic understanding to jointly infer gaze direction and target semantics, (ii) extend gaze following to multi-person scenes without compromising efficiency, allowing real-time deployment, and (iii) model uncertainty explicitly by separating inherent ambiguity in gaze from epistemic uncertainty in model reasoning, resulting in more reliable and interpretable predictions.

The intellectual trajectory of this work follows a natural progression—from pixels to attention. It begins with geometric grounding, then moves toward semantic and social reasoning. Along the way, the thesis reconsiders both the practical utility and ethical framing of gaze prediction, emphasizing accuracy, robustness, and responsible use in socially sensitive domains.

Beyond technical advances, this research contributes a conceptual shift: gaze following is not merely a visual task but a step toward socially intelligent perception. By modeling attention as a relational construct linking vision, semantics, and social inference, it moves gaze understanding closer to the human cognitive process it seeks to emulate. From pixels to attention, it lays the foundations for vision systems capable of perceiving, reasoning, and ultimately engaging with humans in the shared spaces of the real world.

Abstract

Keywords: Gaze Following; Social Gaze Behavior; Multimodal Learning; 3D Scene Geometry; Semantic Grounding; Uncertainty Estimation; Social AI

Résumé

Le regard constitue l'une des fenêtres les plus révélatrices, mais aussi les moins comprises, de la cognition humaine. L'endroit où les personnes dirigent leur regard — vers un objet ou une autre personne — reflète l'attention, la coordination et l'interaction. En vision par ordinateur, cependant, le regard a longtemps été réduit à un simple problème de régression géométrique : prédire un point sur le plan de l'image. Cette formulation saisit la direction spatiale, mais néglige la portée plus large du regard en tant qu'indice pour comprendre le comportement. Ce travail vise à rendre la modélisation du regard plus informative et exploitable en améliorant la précision, la scalabilité et l'interprétabilité grâce à une modélisation plus riche de la géométrie, de la sémantique et de l'incertitude.

Cette thèse redéfinit le suivi du regard comme un pont entre la vision et la compréhension sociale. Elle développe des modèles computationnels qui raisonnent non seulement sur la localisation de la cible du regard, mais aussi sur les relations attentionnelles reliant les personnes, les objets et leur environnement. À travers les dimensions des données, de la géométrie, de la sémantique et de l'incertitude, elle propose une vision unifiée du regard comme un problème de raisonnement structuré inscrit dans des scènes naturelles complexes.

Au niveau des données, nous introduisons de nouveaux jeux de référence capturant plusieurs couches de contexte visuel et social dans des environnements naturels. Au niveau de la modélisation, nous proposons plusieurs avancées architecturales : (i) coupler la géométrie spatiale à la compréhension sémantique pour inférer conjointement la direction du regard et la sémantique de la cible, (ii) étendre le suivi du regard à des scènes multi-personnes sans compromettre l'efficacité, permettant un déploiement en temps réel, et (iii) modéliser explicitement l'incertitude en séparant l'ambiguïté inhérente du regard de l'incertitude épistémique du raisonnement du modèle, conduisant à des prédictions plus fiables et interprétables.

La trajectoire intellectuelle de ce travail suit une progression naturelle — des pixels à l'attention. Elle commence par un ancrage géométrique, puis évolue vers un raisonnement sémantique et social. En parallèle, la thèse réexamine l'utilité pratique et le cadre éthique de la prédiction du regard, en mettant l'accent sur la précision, la robustesse et une utilisation responsable dans des contextes socialement sensibles.

Au-delà des avancées techniques, cette recherche introduit un changement conceptuel : le

suivi du regard n'est pas seulement une tâche visuelle, mais une étape vers une perception socialement intelligente. En modélisant l'attention comme une construction relationnelle reliant vision, sémantique et inférence sociale, elle rapproche la compréhension du regard du processus cognitif humain qu'elle cherche à reproduire. Des pixels à l'attention, elle jette les bases de systèmes de vision capables de percevoir, raisonner et, en fin de compte, interagir avec les humains dans les espaces partagés du monde réel.

Contents

Acknowledgements i

Abstract (English/Français) iii

I Foundations 1

1 Introduction 3
- 1.1 Introduction to Gaze and Attention . . . 3
- 1.2 Motivation and Problem Statement . . . 4
- 1.3 Landscape of the Problem Space . . . 6
- 1.4 Key Contributions . . . 8
- 1.5 Roadmap . . . 8

2 Background and Related Work 11
- 2.1 Core Concepts and Terminology . . . 11
- 2.2 Task Definitions in Computer Vision . . . 13
  - 2.2.1 Gaze Estimation . . . 13
  - 2.2.2 Visual Focus of Attention Estimation . . . 13
  - 2.2.3 Gaze Target Prediction . . . 14
  - 2.2.4 Mutual Gaze Detection . . . 15
  - 2.2.5 Shared Attention Prediction . . . 15
- 2.3 Existing Literature . . . 16
  - 2.3.1 VFOA Estimation . . . 16
  - 2.3.2 Gaze Target Prediction . . . 18
  - 2.3.3 Social Gaze Prediction . . . 23

3 Datasets & Benchmarks 27
- 3.1 Gaze Target Prediction . . . 27
  - 3.1.1 GazeFollow . . . 27
  - 3.1.2 VideoAttentionTarget . . . 30
  - 3.1.3 DLGaze . . . 31
  - 3.1.4 ChildPlay . . . 32
  - 3.1.5 GazeHOI . . . 37
- 3.2 Social Gaze Prediction . . . 40

3.2.1 UCO-LAEO . . . 40
3.2.2 VideoCoAtt . . . 40
3.2.3 VACATION . . . 41

4 Experimental Protocol **45**
4.1 Performance Metrics . . . 45
4.1.1 Distance . . . 46
4.1.2 Area Under the Curve (AUC) . . . 47
4.1.3 Average Precision (AP) . . . 48
4.1.4 Precision of LAH (P_{LAH}) . . . 48
4.1.5 Object Gaze Accuracy (OGA) . . . 48
4.1.6 Semantic Gaze Accuracy (SGA) . . . 49
4.2 Implementation Details . . . 49
4.2.1 Resolution . . . 49
4.2.2 Coordinate Encoding . . . 49
4.2.3 Coordinate Decoding . . . 50
4.2.4 Data Augmentation . . . 51
4.2.5 Context People . . . 51
4.2.6 Loss Functions . . . 52
4.2.7 Training Hyperparameters . . . 53
4.3 Training and Evaluation Routines . . . 54
4.4 On the Limitation of Video-based Benchmarks . . . 55
4.4.1 Qualitative Analysis . . . 55
4.4.2 Quantitative Analysis . . . 55

II Methods **59**

5 Multimodality & Scene Geometry **61**
5.1 Motivation . . . 61
5.2 Related Work . . . 63
5.2.1 Multimodal Fusion . . . 63
5.2.2 Geometric Reasoning . . . 63
5.3 Architecture . . . 65
5.3.1 Gaze Encoder . . . 65
5.3.2 3D Field-of-View Representation . . . 67
5.3.3 Scene Encoder . . . 69
5.3.4 Gaze Decoder . . . 70
5.4 Implementation Details . . . 71
5.5 Experimental Design . . . 72
5.5.1 Benchmark Comparison . . . 72
5.5.2 Ablation Studies . . . 73
5.5.3 Stability of 3D Gaze Vector Supervision . . . 74

5.5.4 Qualitative Evaluation . . . 74
5.6 Limitations . . . 75
5.7 Conclusion . . . 76

6 Efficiency & Multi-Person Scaling **77**
6.1 Motivation . . . 77
6.2 Related Work . . . 79
6.2.1 Multi-Person Gaze Following . . . 79
6.2.2 Efficiency and Decoding . . . 79
6.3 Architecture . . . 80
6.3.1 Scene Tokens . . . 80
6.3.2 Gaze Tokens . . . 80
6.3.3 Scene Encoder . . . 82
6.3.4 Gaze Decoder . . . 82
6.3.5 In/Out Decoder . . . 83
6.4 Implementation Details . . . 83
6.5 Experimental Design . . . 84
6.5.1 Benchmark Comparison . . . 84
6.5.2 Ablation Studies . . . 85
6.5.3 Efficiency Analysis . . . 86
6.5.4 Role of Training Group Size . . . 87
6.5.5 Output Representation . . . 88
6.5.6 Generalization and Applications . . . 91
6.5.7 Qualitative Evaluation . . . 92
6.6 Limitations . . . 92
6.7 Conclusion . . . 94

7 Social Gaze Behavior **95**
7.1 Motivation . . . 96
7.2 Related Work . . . 97
7.2.1 Task-Specific Formulation . . . 97
7.2.2 Unified Formulation . . . 97
7.2.3 Graph Neural Networks . . . 98
7.3 Architecture . . . 98
7.3.1 Gaze Following Core . . . 99
7.3.2 Social Gaze Extension . . . 100
7.4 Implementation Details . . . 101
7.5 Experimental Design . . . 103
7.5.1 Evaluation Protocol . . . 103
7.5.2 Reference Baselines . . . 104
7.5.3 Benchmark Comparison . . . 104
7.5.4 Qualitative Evaluation . . . 106
7.6 Limitations . . . 107

7.7 Conclusion . . . 109

8 Semantic Grounding 111
8.1 Motivation . . . 111
8.2 Related Work . . . 112
8.2.1 Semantic Gaze Following . . . 113
8.2.2 Human-Object Interaction . . . 113
8.2.3 Vision-Language Modeling . . . 113
8.3 Architecture . . . 114
8.3.1 Scene Encoder . . . 115
8.3.2 Gaze Encoder . . . 115
8.3.3 Gaze Decoder . . . 116
8.3.4 Text Encoder . . . 117
8.4 Implementation Details . . . 118
8.5 Experimental Design . . . 119
8.5.1 Reference Baselines . . . 119
8.5.2 Benchmark Comparison . . . 120
8.5.3 Ablation Studies . . . 122
8.5.4 Scene Encoding as Saliency Representation . . . 125
8.5.5 Role of the Background Token . . . 125
8.5.6 Revisiting Efficiency . . . 126
8.5.7 Qualitative Evaluation . . . 127
8.6 Limitations . . . 129
8.7 Conclusion . . . 129

9 Predictive Uncertainty 131
9.1 Motivation . . . 132
9.1.1 Advantages of a Probabilistic Formulation . . . 132
9.1.2 From Point Supervision to Distributional Prediction . . . 133
9.2 Related Work . . . 135
9.2.1 Uncertainty in Deep Learning . . . 135
9.2.2 Distributional and Multi-Modal Regression . . . 136
9.2.3 Alternative Distributional Models . . . 137
9.3 Architecture . . . 138
9.3.1 3D Field-of-View Conditioning . . . 138
9.3.2 Probabilistic Decoder . . . 139
9.3.3 In/Out Prediction . . . 140
9.3.4 Per-Mode Semantic Embeddings . . . 140
9.4 Implementation Details . . . 141
9.5 Experimental Design . . . 143
9.5.1 Benchmark Comparison . . . 143
9.5.2 Ablation studies . . . 143
9.5.3 Uncertainty Calibration . . . 146

9.5.4 Multi-Hypothesis Semantic Embeddings . . . 147
9.5.5 Qualitative Evaluation . . . 148
9.6 Limitations . . . 149
9.7 Conclusion . . . 151

III Discussion & Outlook **153**

10 Limitations & Ethical Considerations **155**
10.1 Methodological Limitations . . . 155
10.1.1 Data-related Limitations . . . 155
10.1.2 Model-related Limitations . . . 156
10.1.3 Evaluation-related Limitations . . . 157
10.2 Ethical & Privacy Considerations . . . 159
10.2.1 Handling Personally Identifiable Visual Data . . . 159
10.2.2 Risk *vs.* Utility . . . 159
10.2.3 Toward Privacy-Preserving Design Principles . . . 160
10.3 Broader Limitations of Gaze Following as a Task . . . 161

11 Open Challenges and Future Directions **165**
11.1 Building Richer Datasets . . . 165
11.2 Leveraging VLMs for Contextual Gaze Reasoning . . . 166
11.3 Temporal Reasoning . . . 167
11.4 Native 3D Gaze Following . . . 168
11.5 Beyond Gaze: Joint Modeling of Social and Motor Behavior . . . 169
11.6 Applications and Societal Impact . . . 171

12 Conclusion **173**

Bibliography **191**

Foundations Part I

1 Introduction

1.1 Introduction to Gaze and Attention

Gaze is a fundamental non-verbal cue that conveys where a person directs their visual focus. It plays a central role in human communication, serving functions such as regulating turn-taking (Degutyte and Astell, 2021), signaling intentions (Emery, 2000), and establishing joint attention (Stephenson et al., 2021). The importance of gaze has been recognized since antiquity, where classical philosophers speculated about the role of the eyes in perception and influence. Plato proposed that vision operated through rays emitted by the eyes (Timaeus), an idea later challenged by Aristotle (De Anima) and definitively refuted by medieval scholars such as Ibn al-Haytham in his Book of Optics (Archer-Hind et al., 1888; Shields et al., 2016; al Haytham, 1989; Lindberg, 1976). This long-standing fascination underscores how deeply eye gaze has been linked to knowledge, power, and interpersonal connection across cultures and history.

In developmental psychology, the ability to follow another's gaze is considered a precursor to the theory of mind (Calder et al., 2002) and is linked to cognitive and language development (Brooks and Meltzoff, 2005). Abnormal gaze patterns are also known markers of neurodevelopmental disorders such as autism (Neumann et al., 2006).

In cognitive science, gaze is closely related but not identical to attention. Attention refers to the allocation of mental processing resources to a specific task, which may or may not align with the physical direction of gaze. For example, a person might keep their eyes on the TV while mentally focusing on a nearby conversation. In other words, gaze can be directly observed as a behavioral signal through eye and head orientation, whereas attention is a latent cognitive process that can only be inferred. Consequently, gaze is often used as a proxy for attention in both behavioral studies and computational models (Bayliss et al., 2011).

From a computer vision perspective, this distinction motivates different tasks: gaze estimation (inferring gaze direction from the eyes or head), visual focus of attention (estimating the gaze target in 3D settings), and gaze following (predicting the 2D gaze target in natural scenes). Together, these tasks contribute to a broader understanding of attention and social behavior,

enabling applications in human-computer interaction, robotics, and medical diagnosis.

Beyond these direct applications, gaze and attention also play a central role in what has been termed the *attention economy* (Hogan, 2001): the modern reality in which human attention is treated as one of the most valuable resources. Social media platforms, streaming services, and advertisers compete to capture and monetize user attention, often optimizing content to maximize "time on screen". In this context, gaze serves not only as a cognitive signal, but also as an economic currency. The advertising industry has long relied on eye-tracking studies to measure engagement, and more recently, cultural critiques have extended the concept to phenomena such as the "male gaze economy", exemplified by marketing campaigns that explicitly exploit how gaze shapes perception and consumer desire. These broader societal developments highlight the enduring relevance of gaze beyond cognitive science and computer vision, situating it at the heart of both technological innovation and cultural discourse.

1.2 Motivation and Problem Statement

Understanding where people look, and by extension, what they attend to, offers a powerful window into the mechanisms of human cognition and communication. Gaze is not merely a visual signal; it is a fundamental expression of thought, intention, and social engagement. From infancy, humans use gaze to coordinate with others, establish joint attention, and infer mental states. These subtle dynamics of looking and being looked at form the substrate of social intelligence, enabling humans to interact naturally in shared environments. Translating this capacity into machines is a long-standing ambition in computer vision and artificial intelligence, with implications ranging from social robotics and education to healthcare and behavioral sciences.

In computer vision, this ambition is captured by the task of gaze following: predicting the location in an image where a person is looking. Unlike gaze estimation, which relies on close-up views of the eyes and face under controlled conditions, gaze following operates from an external, third-person viewpoint in natural environments. This exo-centric formulation makes it particularly appealing for analyzing human interactions in everyday settings, where wearable sensors are impractical and multiple people coexist in complex visual scenes. Yet, it also makes the task profoundly challenging: the model must simultaneously interpret the geometry of the scene, the orientation of the person's head and body, and the contextual cues that signal the focus of attention.

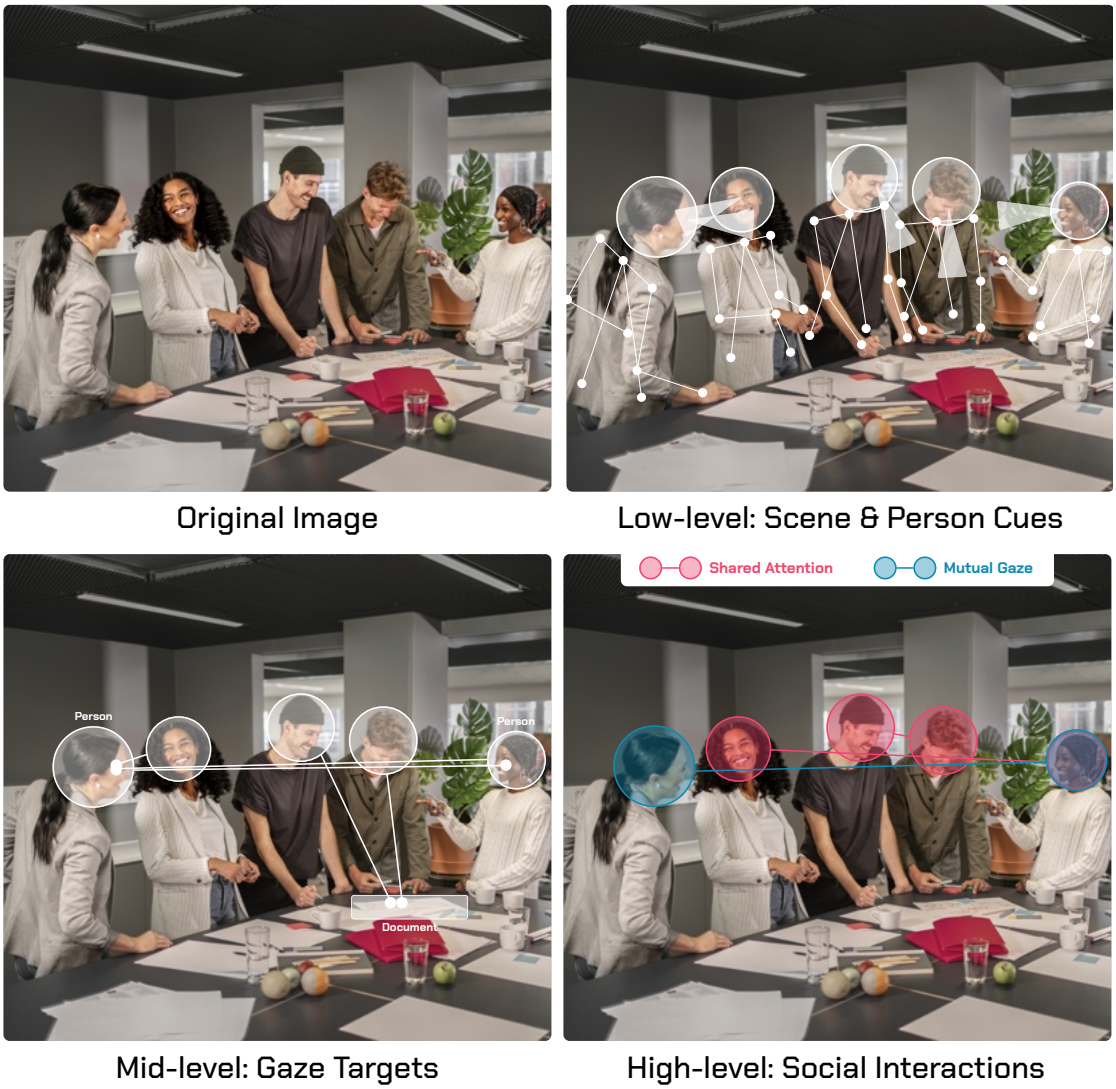


Figure 1.1: From perception to social understanding. Our ability to infer gaze and attention relies on a hierarchy of reasoning steps. Starting from the raw visual scene (top left), models first extract low-level cues describing the 3D scene structure and person-specific signals such as head location, gaze direction, and body orientation (top right). These cues are then combined to infer mid-level gaze targets: the specific locations and underlying objects or people being attended to (bottom left). Finally, aggregating these individual predictions reveals high-level social interactions, such as shared attention and mutual gaze, that underpin joint engagement and communication (bottom right).

Despite notable progress over the past decade, current approaches remain limited in both scope and interpretability. First, most methods are grounded in 2D spatial correlations and overlook the geometric structure of the scene, which is essential for distinguishing between plausible and implausible gaze targets in three dimensions. Second, models trained on existing benchmarks, often composed almost exclusively of adults in scripted or entertainment media, generalize poorly to natural settings, especially those involving children or unstructured social behavior. Third, standard architectures predict gaze for one person at a time, requiring multiple forward passes for multi-person scenes and failing to account for social dependencies

between individuals. Fourth, traditional formulations provide no semantic understanding of what is being looked at; they predict a point on the image plane, not an interpretable entity. Finally, gaze predictions are inherently uncertain, yet existing models rarely quantify or interpret this uncertainty, limiting their usefulness in downstream behavioral analysis.

Taken together, these limitations reveal a deeper issue: the field has largely treated gaze prediction as a spatial localization problem, when in reality it is a reasoning problem: one that demands a joint understanding of geometry, social context, and semantics. Achieving human-like gaze understanding requires models that not only predict *where* people look, but also infer *why* they look there, grounded in the structure and meaning of the visual scene. This calls for moving beyond single-modality architectures toward multimodal, geometry-aware, and semantically grounded frameworks that reflect the richness of human attention in the wild. Figure 1.1 illustrates the progression of reasoning in gaze inference, from low-level perception to high-level social understanding.

Research Question

How can we design computational models that perceive and reason about human gaze in natural exo-centric scenes in a way that captures the spatial, social, and semantic principles that underlie human attention?

This question guides the investigation throughout the dissertation: from geometrically grounded modeling (addressing the spatial component), to multi-person and social reasoning (capturing the social dimension), to semantic and probabilistic formulations (addressing meaning and uncertainty). Each successive chapter expands the scope of gaze understanding toward the broader goal of building computational systems capable of reasoning about human attention in natural environments.

1.3 Landscape of the Problem Space

Human gaze is a complex signal that integrates perception, cognition, and social intent. Therefore, modeling it computationally spans multiple dimensions, each revealing a different facet of human attention. The research field of gaze following has evolved around several core challenges and open directions, which collectively define the broader landscape of the problem space.

Datasets. Progress in gaze understanding is tightly coupled with the quality and diversity of datasets. Existing benchmarks often feature low-resolution imagery, narrow contextual coverage, and limited gaze classes, typically restricted to “inside” or “outside” the frame. Future datasets must expand these taxonomies to include richer gaze classes (e.g. gaze shifts, closed eyes, occlusion flags) and higher-quality, object-centric annotations with bounding boxes, segmentation masks, and semantic labels.

Evaluation Metrics. Standard metrics such as pixel-wise distance measure spatial accuracy but overlook the semantic and social meaning of gaze. A gaze prediction landing a few pixels outside a person's head or object can be spatially accurate yet semantically incorrect. There is thus a need to complement dense metrics with object-centric and relation-centric evaluations that measure whether the predicted gaze aligns with the intended entity or social interaction.

Spatial Reasoning. At the most fundamental level, gaze is a geometric process: it depends on the spatial configuration between the observer and their surroundings. Inferring where people look thus requires reasoning about head orientation, depth, occlusions, and 3D scene structure. Most early models operated purely in 2D, ignoring geometry and resulting in physically implausible predictions. Recent work moves toward explicit 3D representations and field-of-view modeling, a step toward more interpretable and geometrically consistent gaze reasoning.

Social Reasoning. Gaze is fundamentally a social signal. In multi-person scenes, people's attention is often interdependent: directed at, coordinated with, or shared among others. Understanding such interactions requires models that go beyond individual gaze estimation to infer social behaviors such as looking at others, mutual gaze, or shared attention. These social dimensions are central to human communication and underlie the emergence of joint behavior and coordination.

Semantic Reasoning. Beyond spatial localization, a complete account of gaze must capture *what* entities are being attended to. Semantic gaze prediction links visual perception to conceptual understanding, grounding gaze in the meaning of objects and scene context. It enables models to reason about attention in a goal-directed manner (e.g. identifying that a person is looking at a phone, a toy, or another human), rather than predicting undifferentiated image coordinates.

Temporal Reasoning. Attention is inherently dynamic, unfolding over time through saccades, fixations, and gaze shifts. Temporal reasoning is crucial for modeling turn-taking, engagement, and gaze transitions in social exchanges. While most existing datasets and models treat gaze statically, temporal modeling remains an open research frontier explored in this thesis mainly through future outlook and discussion.

Efficiency and Scalability. Most current approaches operate under a single-person formulation: predicting one person's gaze per forward pass. This constraint severely limits deployment in real-time or multi-agent scenarios where many individuals must be analyzed simultaneously. Efficient, transformer-based, and tokenized representations enable parallel reasoning over all people in a scene, a key step toward scalable and deployable systems.

Uncertainty Modeling. Human attention is probabilistic and context-dependent. Ambiguity arises from limited visual evidence (e.g. occluded eyes), multiple plausible gaze targets, or inherent subjectivity in annotation. Representing this uncertainty explicitly allows models

to provide calibrated predictions, avoid overconfidence, and communicate ambiguity in human-interpretable ways that simplify decisions in downstream applications.

Gaze following sits at the intersection of cognitive science, computer vision, and artificial intelligence. It draws from cognitive models of attention and joint intention, connects to multimodal and vision-language modeling through semantic grounding, and intersects with embodied AI and robotics where gaze provides a natural interface for human–machine interaction. This interdisciplinary overlap makes gaze a uniquely integrative research problem: both perceptual and social, both physical and semantic.

The present thesis traverses these dimensions progressively. It begins with the creation of new datasets and evaluation protocols, then develops models that address geometric reasoning, scalability and multi-person interaction, semantic grounding and uncertainty, and finally looks forward to temporal and contextual extensions. These stages outline a coherent progression from low-level spatial perception to high-level, interpretable, socially-aware, and practically valuable gaze understanding.

1.4 Key Contributions

This section summarizes my individual contributions across the main publications produced during this doctoral work (*cf.* Table 1.1). Each project involved regular discussions with my advisor and labmates, with reciprocal participation in each other's research meetings. Roles are defined as follows:

- **Lead Investigator:** led the study from conception to writing
- **Joint Lead Investigator:** shared equal responsibility with a co-author
- **Contributing Investigator:** provided targeted technical or conceptual contributions.

1.5 Roadmap

This thesis is organized into three parts, with the remaining chapters progressing from foundational principles and datasets to modeling approaches and finally to broader reflections and future directions.

Part I – Foundations introduces the conceptual and methodological basis for this work.

- **Chapter 2** reviews the relevant literature on gaze estimation, gaze following, and related visual attention tasks, situating this research within the broader landscape of human-centric computer vision.
- **Chapter 3** presents the datasets and benchmarks used throughout the thesis, including large-scale public datasets such as *GazeFollow* and *VideoAttentionTarget*, as well as the newly proposed *ChildPlay* dataset for child gaze behavior and *GazeHOI* for object-centric gaze understanding.

Publication	Venue	Role	Summary of Contributions
Gupta et al. (2022)	*CVPRW*	Contributor	Pre-processed data and modalities, trained the head-centric branch, conducted experiments with the full architecture, co-wrote the paper, and created the main figure.
Tafasca et al. (2023b)	*ICCV*	Joint Lead	Curated and processed videos for the ChildPlay dataset, drafted the annotation protocol, recruited and supervised annotators, performed quality control and managed finances. Conducted part of the experiments, co-wrote the paper, and produced all figures.
Tafasca et al. (2023a)	*ICMIW*	Lead	Initiated and coordinated the paper summarizing the outcomes of the AI4Autism project. Led the writing effort, integrated partner contributions, and structured the paper.
Tafasca et al. (2024b)	*CVPR*	Lead	Conceived the idea and architecture, implemented and trained the model, conducted all experiments and ablations, analyzed results, and wrote the entire paper including all figures.
Gupta et al. (2024a)	*FG*	Joint Lead	Proposed the original idea of extending *Sharingan* to social gaze using a graph-based interaction between gaze tokens. Shared implementation, experimentation, and analysis; co-wrote the paper and created the figures.
Gupta et al. (2024b)	*NeurIPS*	Contributor	Provided the conceptual inspiration for the architecture (ViT-Adapter–inspired design). Contributed the gaze decoder based on *Sharingan* and created the main architecture figure.
Tafasca et al. (2024a)	*NeurIPS*	Lead	Proposed and implemented the semantic gaze formulation, developed the multimodal architecture, collected and annotated the GazeHOI dataset, generated semantic pseudo-labels for GazeFollow, conducted all experiments, and wrote the paper.
Farkhondeh et al. (2024)	*ECCVW*	Joint Lead	Conceived and created the ChildPlay-Hand dataset, drafted the annotation protocol and contracts, recruited and supervised annotators, and guided the experimental design. Contributed to writing, analysis, and figure production.
ProbGaze	–	Lead	Designed the probabilistic gaze model with mixture-density decoding, implemented the framework, performed experiments and uncertainty analyses, and authored the entire paper.

Table 1.1: Summary of my roles, venues, and specific contributions across the main publications produced during this thesis.

- **Chapter 4** describes the experimental protocol, including data preprocessing, evaluation metrics, and cross-dataset training strategies, establishing a consistent framework for subsequent experiments.

Part II – Methods develops a sequence of modeling advances that progressively expand the scope of gaze understanding.

- **Chapter 5** introduces the first core methodological contribution: a **multimodal and geometry-aware** approach to gaze target prediction that explicitly models the 3D structure of the scene.
- **Chapter 6** presents a transformer architecture designed for efficient and scalable **multi-person gaze following**, capable of predicting gaze for all individuals in a scene simultaneously.
- **Chapter 7** extends this framework to model **social gaze behaviors**, such as looking at humans, mutual gaze, and shared attention, introducing a unified multi-task formulation.
- **Chapter 8** explores **semantic grounding**, proposing a multimodal approach that predicts not only *where* a person is looking but also *what* they are attending to, through joint visual–language representations.
- **Chapter 9** addresses **predictive uncertainty**, reformulating gaze prediction as a probabilistic inference problem to quantify ambiguity and confidence in model predictions.

Part III – Discussion and Outlook reflects on the broader implications of this work, addressing limitations, ethical considerations, and future perspectives.

- **Chapter 10** discusses methodological and ethical limitations of current approaches, including dataset bias, annotation quality, and the implications of modeling human behavior.
- **Chapter 11** outlines open challenges and future research directions, highlighting promising application domains such as social robotics, autism assessment, and education analytics.
- **Chapter 12** concludes the thesis.

Together, these chapters follow a coherent trajectory: from geometric perception to social and semantic reasoning, culminating in a comprehensive framework for computational gaze understanding in natural, exo-centric scenes.

2 Background and Related Work

In this chapter, we define gaze in relation to cognitive concepts such as attention, provide an overview of the existing terminology, introduce the main gaze-related tasks in computer vision and machine learning, and conclude with a review of existing research in the field.

2.1 Core Concepts and Terminology

The field of gaze research combines perspectives from cognitive science, psychology, and computer vision, which has led to overlapping yet sometimes inconsistent terminology. To establish clarity, we briefly define the core concepts that will be used throughout this thesis, and summarize them in Table 2.1.

At the most fundamental level, *gaze* refers to the direction of one's eyes (and by extension, the head), providing an explicit and observable signal of where someone is looking. In contrast, *attention* denotes the internal allocation of cognitive resources toward a stimulus or task. As explained before, while gaze and attention are often aligned, they can also diverge. To emphasize this distinction, *gaze* is observable, whereas *attention* is latent.

Within this broader framework, *visual attention* designates the subset of attention directed toward visual stimuli, which may be overt (accompanied by gaze shifts) or covert (without eye movements). *Shared attention* and *joint attention*, while often used interchangeably, describe different higher-order social behaviors: the former refers to the situation where two or more individuals focus on the same object or event (e.g. people watching a soccer game in a pub), while the latter is a more complex, reciprocal form of shared attention that involves coordination and mutual awareness. In other words, participants know that they are attending to the same thing and have the intent to engage with one another about it (e.g. a mother and her child looking at the toy they are playing with). Finally, *mutual gaze*, or *eye contact*, is the special case where two individuals look directly at each other, a key signal in social interaction. The phrase *eye contact* is also used to describe the computer vision task of predicting whether a person is looking at the camera in egocentric settings.

On the methodological side, *eye-tracking* refers to hardware-based techniques for measuring eye movements at fine temporal and spatial resolution, enabling detailed analyses of saccades and fixations. However, such systems require specialized, often intrusive equipment and are not practical for unconstrained, in-the-wild settings. To overcome these limitations, computer vision methods focus on *gaze estimation*, which aims to infer gaze direction (2D or 3D vector) directly from visual inputs such as the eyes, face, or head.

Gaze following, which is central to this thesis, is a broader human ability: the capacity to infer another person's focus of attention based on gaze cues and the surrounding context, such as the spatial layout of the scene, the presence of objects, and the actions of other people. In computer vision and machine learning, this is formalized as *gaze target prediction*: given an image and the location of a person, the goal is to predict the 2D position of the gaze target in natural images or videos without requiring additional equipment.

Together, these definitions delineate the landscape of low-level perceptual signals (gaze, eye-tracking), intermediate computational tasks (gaze estimation, gaze target prediction), and higher-level social constructs (mutual gaze, shared attention).

Concept	Definition
Gaze	Observable direction of the eyes (and head), indicating where a person is looking.
Attention	Latent cognitive process of allocating mental resources; not always aligned with gaze.
Visual Attention	Subset of attention directed to visual stimuli; can be overt (with gaze) or covert (without gaze).
Saliency	Visual distinctiveness of scene regions that makes them likely to attract gaze.
Shared Attention	Multiple individuals attending to the same target without engagement.
Joint Attention	Coordinated focus of two or more individuals on the same target, with mutual awareness and active engagement toward that object or event.
Mutual Gaze / Eye Contact	Two individuals looking directly at each other; a strong social cue.
Eye-Tracking	Hardware-based measurement of eye movements (fixations, saccades, etc.) with high precision, using a head-mounted device.
Gaze Estimation	Computer vision task of predicting gaze direction (2D/3D vector) from visual inputs such as eyes or face.
Gaze Following / Gaze Target Prediction	Human ability to infer others' focus of attention from gaze cues and scene/social context; in ML, the task of predicting the 2D gaze target in images or videos.

Table 2.1: Core concepts and terminology used in gaze and attention research.

2.2 Task Definitions in Computer Vision

In this section, we introduce the core computer vision tasks that have been proposed to model gaze and its social manifestations. These task definitions provide a common ground for comparing methods, datasets, and evaluation protocols, and help clarify the distinctions between related problems such as gaze estimation, gaze following, mutual gaze, and shared attention. By outlining each task, we highlight not only their specific input–output formulations, but also their role in advancing machine perception of human attention and interaction. We also provide a summary of these tasks in Figure 2.1 to illustrate their inner workings and differences.

2.2.1 Gaze Estimation

Gaze estimation is one of the earliest tasks proposed in gaze analysis, and aims to infer a person's gaze direction directly from visual input, thereby bypassing the need for intrusive head-mounted eye-trackers. Typically, the input is an image of the head or face crop $\mathbf{I}_{\text{head}}$, from which the model predicts a 3D gaze vector $\mathbf{g}_{\text{vec}}$ representing the line of sight. The task leverages both fine-grained ocular cues (eye orientation) and coarse head pose information to estimate where a person is looking.

This problem formulation is well established and has been extensively studied, with methods ranging from geometric models based on head pose to modern deep learning approaches trained on large-scale datasets such as Gaze360, which was introduced in Kellnhofer et al. (2019). As a result, gaze estimation has become a popular benchmark task in computer vision, spawning numerous challenges, datasets, and research directions. Beyond its technical appeal, gaze estimation also serves as a key building block for higher-level tasks such as gaze following, visual focus of attention prediction, and social gaze inference.

2.2.2 Visual Focus of Attention Estimation

Visual Focus of Attention (VFOA) is the task of determining which entity (person or object) a person is attending to in a scene. Classical formulations assume a closed set of candidate targets (e.g., meeting participants, a display, the table) and estimate, for each subject, a discrete target label at each time instant. Inputs typically include tracked head (and sometimes eye) poses, person detections, and scene geometry or calibrated target locations; some systems also leverage audio cues (e.g., current speaker) and conversational context. Outputs are per-person categorical assignments such as "looking at Salma," "looking at the screen," or "elsewhere," often with temporal smoothing via Hidden Markov Models (HMM) or Deep Belief Network-style (DBN) dynamics.

The closed-world assumption, the use of head pose as a gaze proxy, the dependence on camera calibration and the sensitivity to scene context make VFOA accurate and interpretable

in structured environments (e.g. meetings, human-robot interaction demos) but less robust in unconstrained, open-world scenes where targets are not enumerated a priori. This explains why early VFOA systems achieved good performance in fixed environments but struggled to generalize to arbitrary scenes.

2.2.3 Gaze Target Prediction

The task of Gaze Target Prediction or Gaze Following was introduced in Recasens et al. (2015) to generalize gaze understanding in exocentric scenes, and address the limitations of VFOA. Specifically, the goal is to take an image $\mathbf{I}$ and a person's head bounding box $\mathbf{H}_{\text{bbox}}$ as input, and predict the pixel location (x, y) where the person is looking. Since this is a keypoint prediction task, it is commonly framed as a heatmap prediction problem. The idea is to represent the location of each keypoint not as a single pixel, but as a 2D probability distribution spread over the image. This captures the spatial uncertainty around the true position and prevents the model from collapsing its prediction into a single point estimate, which could be misleading or unstable. The point estimate can then be retrieved from the heatmap using the argmax operator. The same idea is used for other keypoint-based tasks like pose estimation and facial keypoints detection.

Alternatively, some works tackle the task directly from the image, jointly detecting head bounding boxes and their corresponding gaze targets. In this formulation, the problem becomes one of detecting paired (person, gaze) instances.

Instead of treating the problem as keypoint prediction, some works have reformulated it as bounding box prediction for the gaze target object . While this approach may appear more practical, it has a critical limitation: it only works for countable objects (commonly referred to as *things* in the literature, e.g., a book), but fails for uncountable regions (*stuff*, such as the ocean or the sky). In addition, datasets supporting this formulation are scarce and highly constrained, which further limits its applicability.

Finally, we can extend the gaze following task by predicting an additional binary flag, which we call *in/out*. This flag indicates whether the person of interest is looking at a target inside or outside the image boundaries. This extension, first introduced in Chong et al. (2020b), adds an important distinction: it allows the model to capture cases where the gaze is directed beyond the visible scene rather than forcing it to choose a point within the image.

Gaze following offers a scalable, non-intrusive way to estimate attention in unconstrained images without requiring scene calibration, multi-view setups, or a closed set of targets. In practice, it underpins downstream applications in social behavior analysis and HRI (attention-aware interaction, turn-taking), video understanding and surveillance (saliency-driven indexing), retail and sports analytics (engagement measurement), and clinical research (early indicators of atypical social attention). It also serves as a foundation for higher-level social gaze tasks such as mutual gaze and shared attention. Much of the work presented in this thesis

will focus on this task and its variations.

2.2.4 Mutual Gaze Detection

Mutual gaze, or equivalently *eye-contact*, is a higher-order social gaze cue that arises when two individuals look directly at each other. In computer vision, the detection of this behavior is typically framed as a binary classification problem, where the input is the scene image $\mathbf{I}$ along with the bounding boxes of two individuals, $\mathbf{H}^1_{\text{bbox}}$ and $\mathbf{H}^2_{\text{bbox}}$, and the output is a prediction indicating whether the pair is engaged in mutual gaze. This task is also commonly termed *Looking At Each Other (LAEO)* in the literature.

A straightforward formulation requires enumerating all N people in the image and testing each of the $\frac{N(N-1)}{2}$ pairs, which quickly becomes computationally expensive in crowded scenes. To address the combinatorial limitation of this approach, recent works such as Guo et al. (2022) have drawn inspiration from object detection, proposing one-stage methods that directly infer mutual gaze instances from the image without explicit pair enumeration. These methods treat mutual gaze as a visual "interaction object", allowing more scalable inference in multi-person scenarios.

Importantly, the phrase *eye-contact detection* is sometimes used in a broader sense to denote whether a person is looking at the camera itself. In this case, the camera is treated as a proxy participant engaged in gaze, and the task reduces to predicting a binary flag from a head or face crop input. This variant is especially relevant in egocentric or first-person vision, where the camera corresponds to the viewpoint of a robot or a VR/AR headset user.

Mutual gaze and eye-contact are fundamental building blocks of social communication: they regulate turn-taking, signal attention, and often precede higher-order behaviors such as joint attention or shared intentionality. Consequently, the automatic detection of mutual gaze has found applications in domains as varied as social robotics, human–computer interaction, behavioral analysis, and developmental psychology (e.g., identifying early markers of atypical social engagement in children).

2.2.5 Shared Attention Prediction

Shared attention is a social gaze phenomenon that occurs when two or more individuals focus their gaze on the same target, thereby establishing a common point of reference. In computer vision, the prediction of shared attention was first introduced by Fan et al. (2018), who framed the task as detecting the bounding box of a shared attention object $\mathbf{SA}_{\text{bbox}}$, given an input scene image $\mathbf{I}$ and the head bounding boxes $\mathbf{H}^i_{\text{bbox}}$ of all visible people. Since a scene may contain zero, one, or multiple shared attention instances, models must be able to handle a variable number of outputs.

Beyond bounding-box localization, shared attention can also be formulated more abstractly

as the detection of non-overlapping sets of people who simultaneously fixate on the same visual target. This relational framing was later adopted in follow-up works such as Fan et al. (2019), where the emphasis is on grouping people based on their gaze alignment rather than identifying a physical object. Recent approaches like Chong et al. (2020b) extend this view further by combining shared attention inference with gaze following predictions: if individual gaze heatmaps converge on the same region or object, a shared attention event can be inferred.

From a behavioral perspective, shared attention is a cornerstone of social cognition: it underpins communication, collaboration, and the development of theory of mind. In developmental psychology, impairments in initiating or responding to shared attention are among the earliest diagnostic markers of Autism Spectrum Disorder (ASD). Consequently, automatic shared attention prediction has practical applications in clinical assessment, human–robot interaction, and the study of group dynamics in naturalistic settings.

2.3 Existing Literature

2.3.1 VFOA Estimation

Definition & Scope. VFOA asks *which discrete target* (e.g. another person, whiteboard, screen, table, "other") a subject attends to in a calibrated social scene, which is typically a meeting room or a constrained HRI setting, making it a structured classification task rather than free-form localization.

Traditional Models. Early systems relied on head orientation as a proxy for attention and mapped it to a small set of 3D target directions, with temporal smoothing via HMM/DBN. Audio features (who is speaking) were often fused to exploit conversational cues. Canonical examples include Stiefelhagen et al. (2001)'s audio-visual meeting setups and Otsuka et al. (2006)'s dynamic Bayesian models.

From individual to joint inference. The work of Ba and Odobez (2010) showed that jointly estimating all participants' VFOA and injecting meeting context (e.g. speaker identity) yields sizable gains over per-person models. Later "collective" formulations (ICAF), proposed in Bai et al. (2019), further leveraged speaking probabilities and temporal dependencies.

Deep and hybrid approaches. Recent works like Zhang et al. (2019) and Otsuka et al. (2018) add CNN features and prosodic signals but still ground predictions in head/eye cues and scene geometry, reflecting the domain's structured nature.

In this manuscript, we build on VFOA to motivate broader *gaze following* and *social gaze* tasks that relax the closed-set assumption by directly localizing, and when needed, semantically labeling the target in unconstrained scenes.

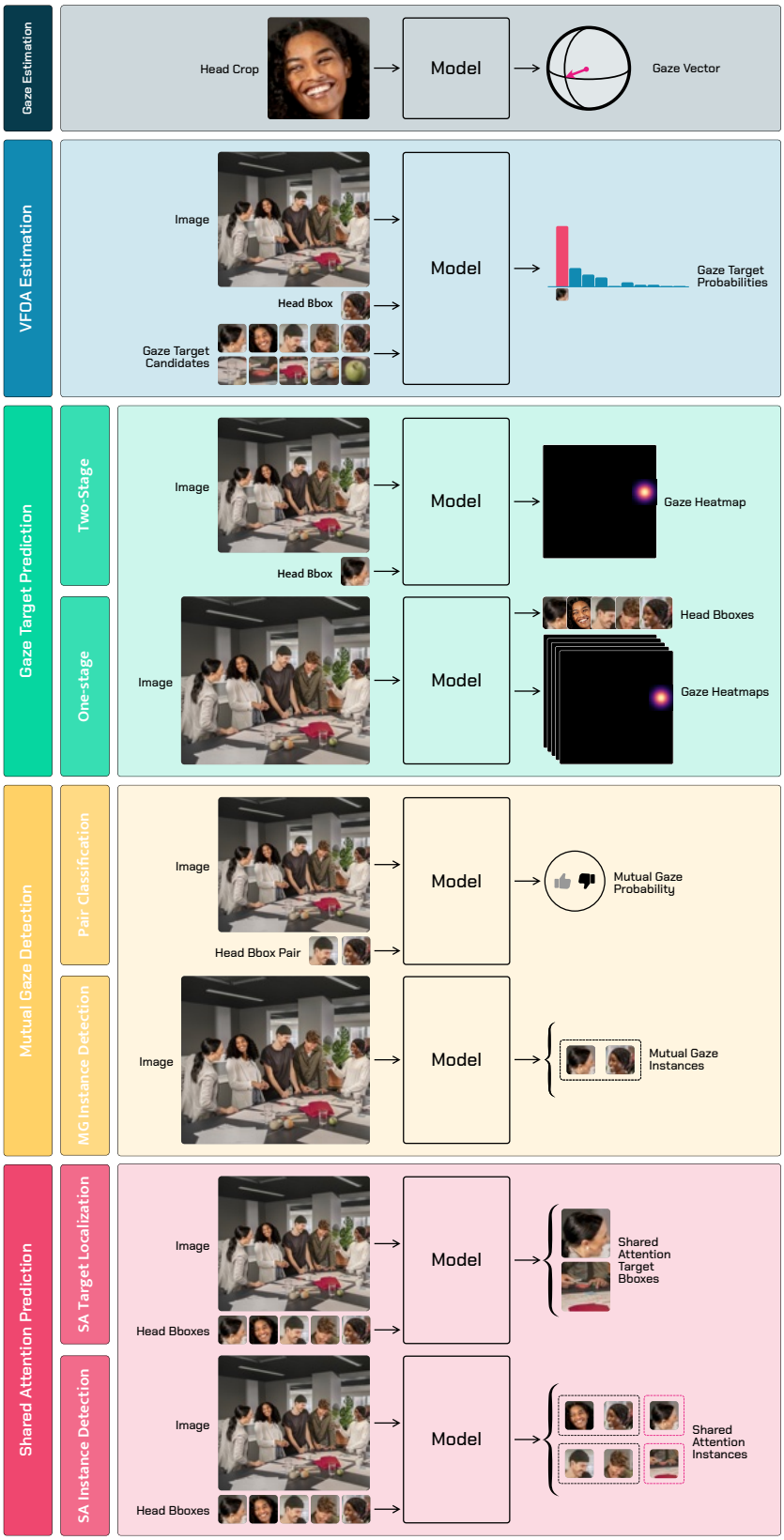


Figure 2.1: Typical formulations of gaze-related tasks in computer vision.

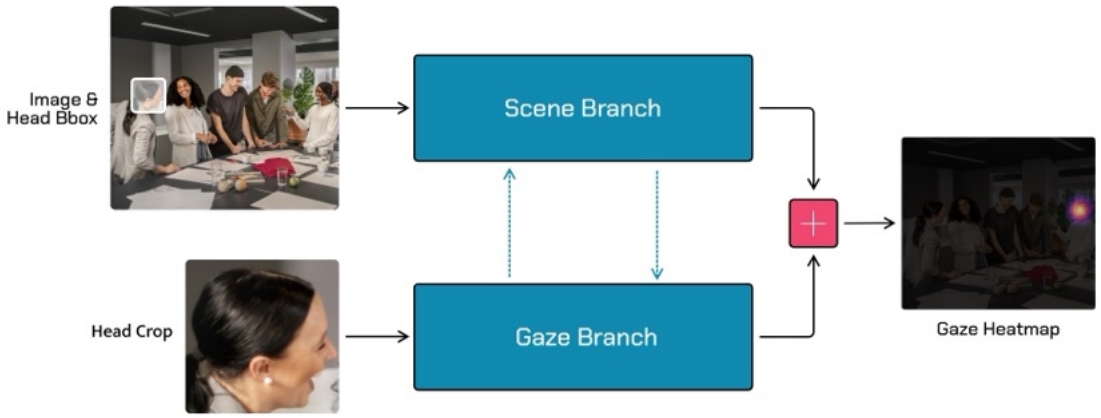


Figure 2.2: Typical two-branch architecture for gaze following. The gaze branch processes the head crop of the person to infer a gaze direction, while the scene branch processes both the input scene image and the head bounding box location (often represented as a binary mask) to capture saliency and context. A fusion mechanism then combines the information to predict the gaze heatmap.

2.3.2 Gaze Target Prediction

Definition & Scope. As explained before, *gaze target prediction*, or equivalently *gaze following*, asks *which pixel location* a subject in the scene attends to. This task applies naturally to unconstrained images and does not require any equipment or prior knowledge about the scene, making it a highly scalable and versatile approach to gaze understanding in-the-wild. To account for cases where a person is gazing outside the image boundaries, we also predict a binary flag to determine when the output gaze point is informative.

Traditional Models. The gaze following task was first introduced in the seminal work of Recasens et al. (2015), which motivated the problem, proposed a CNN-based model, and released a large-scale benchmark for training and evaluation. Since solving the task requires understanding both the subject and the surrounding scene, most subsequent methods adopted a two-stream architecture: one branch processes a high-resolution crop of the head to estimate a gaze direction or field of view, while the other branch analyzes the full scene, often conditioned on the subject's location, to capture salient regions and candidate targets. A fusion module then integrates information from both streams to select the most plausible target and distribute probability mass across competing candidates. We illustrate this canonical design in Figure 2.2.

A notable extension was proposed by Chong et al. (2020b), who introduced the *in/out* prediction task, distinguishing whether the gaze lies inside or outside the image frame. To support this, they collected a new video-based benchmark annotated with both gaze target points and *in/out* labels. Their architecture preserves the two-branch structure but incorporates spatio-temporal modeling for video input, and outputs an *in/out* probability jointly with the gaze heatmap (see Figure 2.3).

Several other works follow this traditional paradigm, refining different components of the pipeline to improve performance; for example, enhancing feature fusion or incorporating eye

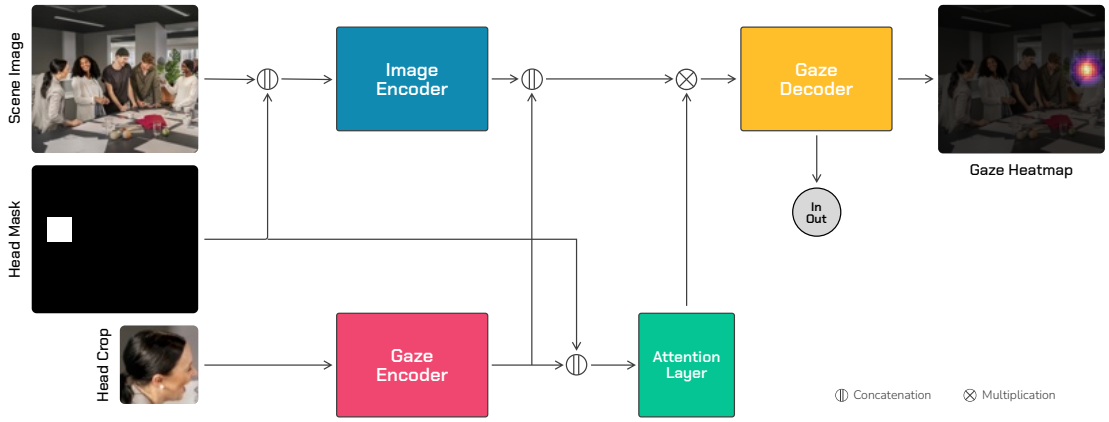


Figure 2.3: Two-stream architecture of Chong et al. (2020b) which encodes the head location as a binary mask, predicts an *in/out* flag and uses a conv-LSTM block for spatio-temporal processing in the gaze decoder when using video clips as input. The figure is adapted from the original paper.

crops (Chong et al., 2018; Wang et al., 2023).

An important distinction to note is that gaze following operates on *heads* rather than strictly on faces. This is because gaze targets can often be inferred, albeit with reduced confidence, from head orientation and scene context, even when the face or eyes are partially occluded or not visible at all. We illustrate such cases in Figure 2.4.

Geometry and Contextual Cues. Gaze following behavior is strongly modulated by cues such as visual saliency and social context (Shepherd, 2010). Humans naturally rely on such priors to resolve ambiguities, drawing on both the 3D spatial layout of the scene and the current context of the subject (e.g. task being performed, social interactions, latent intentions). It is therefore reasonable to assume that predicting a person's gaze target can benefit from explicitly exploiting these cues. While this information is implicitly present in the image, there is no guarantee that a model will learn to capture and integrate it effectively without guidance. For this reason, many approaches have sought to inject geometric and contextual priors as inductive biases into the gaze following process.

One common line of work estimates a 2D gaze vector from the input head and constructs a gaze cone or line of sight to approximate the field of view in the image plane (Lian et al., 2018; Zhao et al., 2020; Fang et al., 2021; Tonini et al., 2023). This narrows down the range of plausible gaze targets before the final selection stage. Another research direction incorporates depth information, typically obtained from off-the-shelf monocular depth estimators, to reason about the 3D structure of the scene (Wei et al., 2018; Fang et al., 2021; Bao et al., 2022; Tonini et al., 2022; Jin et al., 2022; Hu et al., 2022a,b; Tonini et al., 2023; Miao et al., 2023; Hu et al., 2023; Yang and Lu, 2024; Liu et al., 2024; Mathew et al., 2025). Beyond depth, other modalities have also been investigated: body pose (Wei et al., 2018; Guan et al., 2019; Bao et al., 2022; Toaiari et al., 2024; Yang and Lu, 2024), audio or speech cues (Hou et al., 2024), and object or activity features (Wei et al., 2018; Guan et al., 2019; Hu et al., 2022b; Tonini et al., 2023; Yang and Lu, 2024).

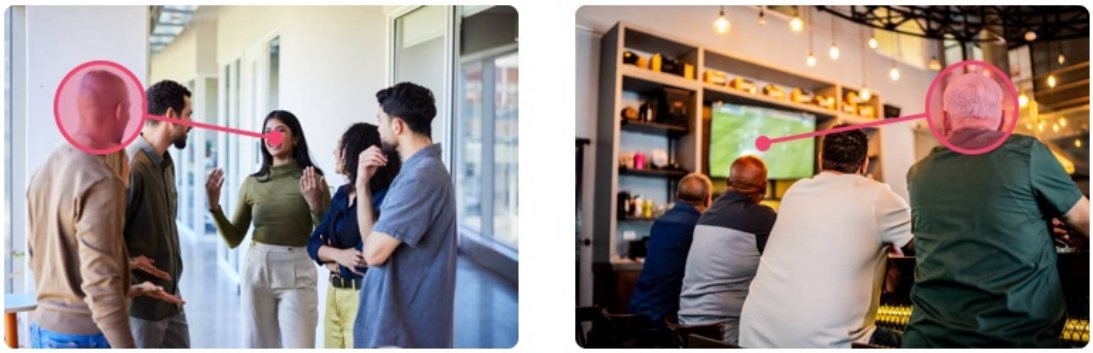


Figure 2.4: Gaze following is possible even when faces are partially or fully occluded. Contextual cues (e.g. main speaker) and head pose may contain sufficient information to produce a reliable prediction.

More recently, Miao et al. (2025) introduced a multi-view gaze following dataset, allowing models to integrate information from multiple synchronized camera viewpoints. This multi-view setting helps disambiguate both subject features and gaze targets, enabling the prediction of targets even when they are visible only from a different viewpoint. This is not the only research to tackle gaze following from an extended scene representation; the work of Li et al. (2021b) for instance, adapted the task to 360-degree imagery. This setting introduces unique challenges and breaks some of the core assumptions of earlier approaches: for instance, the spherical-to-planar projection means that a target is not guaranteed to lie along the 2D line of sight, complicating traditional geometric reasoning.

One-Stage *vs* Two-Stage. Most gaze following methods adopt a *two-stage* paradigm. In this setting, heads are first detected (a step often omitted in figures), and each head is then paired with the scene image to predict the corresponding gaze target. By contrast, Tu et al. (2022) introduced a *one-stage* end-to-end formulation, where the model takes only the image as input and jointly regresses both head bounding boxes and their associated gaze heatmaps. Their architecture follows the DETR framework (Carion et al., 2020), casting gaze target prediction as a set prediction problem in which the output is a collection of *gaze instances*, i.e. pairs of head boxes and gaze targets. We refer the reader to Figure 2.5 which provides an overview of their DETR-inspired architecture. To support this formulation, they also proposed a new evaluation protocol based on bipartite matching at test time. Several follow-up works extended this idea by incorporating additional modalities, modifying the backbone, or refining the decoding stage to improve performance (Tonini et al., 2023; de Belen et al., 2025; Lin et al., 2025). Both *one-stage* and *two-stage* paradigms are illustrated in Figure 2.1.

From a research standpoint, the one-stage formulation is appealing. It is conceptually elegant, unifies head detection and gaze prediction into a single framework, and avoids the need for explicitly conditioning on detected heads. In principle, this efficiency and simplicity could make it more scalable in settings where annotations are dense and exhaustive. Moreover, casting gaze following as "gaze instance detection" opens up avenues for adapting techniques from object detection and set prediction to gaze understanding.

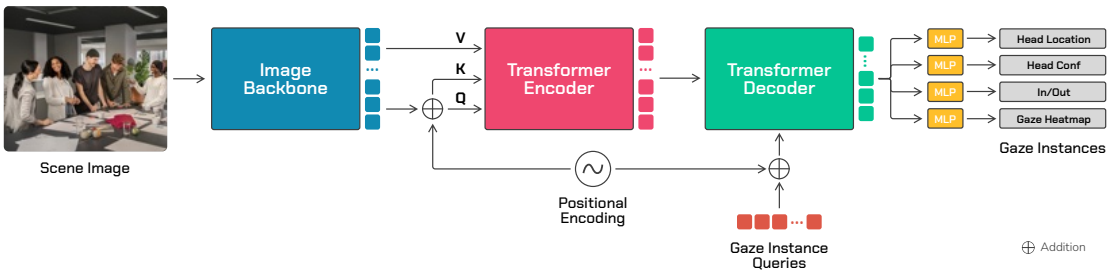


Figure 2.5: Overview of the one-stage architecture from Tu et al. (2022). A convolutional backbone processes the scene image and flattens the spatial feature maps into a sequence of tokens. Positional encoding is added before feeding into a self-attention-based transformer encoder to produce a token-based scene context. Learnable gaze instance queries called HGT are combined with the encoded scene context in a cross-attention-based transformer decoder to produce (head bbox, heatmap, inout, confidence) tuples. The figure is adapted from the original paper.

In practice, however, the two-stage paradigm remains more advantageous in most cases. First, existing benchmarks typically annotate only a subset of people per image (often a single individual). This makes it impossible to fairly assess recall and precision within a detection-style framework, and forces one-stage methods to rely on evaluation protocols with Hungarian matching that depend on ground-truth gaze targets at test time, thereby leaning towards inflated scoring, and limiting comparability with two-stage methods (Ryan et al., 2025). Second, head detection itself is a largely solved problem: modern detectors are highly accurate, robust, and even deployable in real time. Re-learning this step within a gaze network is redundant and diverts model capacity away from the harder problem of gaze understanding. Third, two-stage methods are better aligned with real-world pipelines. In applications such as social robotics or behavioral analysis, people are already detected and tracked as part of broader systems. Conditioning gaze prediction on these known heads simplifies downstream analysis and avoids the additional matching step required in one-stage approaches, which can introduce errors and computational overhead. Finally, DETR-like architectures are known to be slow to train due to the Hungarian matching step (Zhang et al., 2022) and often require high input resolutions (800–1333 pixels) to preserve head detail lost without cropping. As a result, one-stage gaze models tend to be significantly slower than their two-stage counterparts in both training and inference, contradicting the efficiency argument often used to motivate them.

In summary, while one-stage formulations are intellectually interesting and open promising research directions by unifying detection and gaze prediction, the two-stage paradigm is better aligned with the realities of current datasets, benchmarks, and practical applications. For these reasons, most of the architectures presented in this thesis will focus on the two-stage formulation, while recognizing the one-stage line of work as an important research exploration.

From Keypoints to Bounding Boxes. While most gaze following methods cast the task as keypoint detection, or equivalently heatmap regression, Tomas et al. (2021) argued for an object-centric formulation more closely aligned with real-world applications, where the goal



Figure 2.6: An example of annotations in the GOO dataset from Tomas et al. (2021). Can you tell from the RGB image which object the subject is fixating on? The best we can do is narrow the list down to the orange and white products on the second shelf. This is an example of aleatoric uncertainty: there aren't enough person-level and context-level cues for a precise object-level resolution.

is not just to know "where" someone is looking but also "what" they are looking at. To this end, they introduced the GOO dataset, a large-scale collection of images annotated with gaze object bounding boxes, classes, and segmentation masks, consisting of both real and synthetic images (with the synthetic portion covering about 95% of the total data). In practice, however, their work still treated the task as standard gaze following by defining the ground-truth gaze point as the center of the target object's bounding box, rather than directly predicting the bounding box itself.

Building on this direction, Wang et al. (2022) proposed a two-stage architecture that explicitly predicts both the gaze heatmap and the corresponding gaze object bounding box. Subsequent works refined this paradigm by adopting stronger backbones (e.g., transformers in place of CNNs), improving efficiency, or introducing auxiliary supervision such as pixel-level segmentation losses to strengthen object-level localization (Tu et al., 2023; Wang et al., 2024; Jin et al., 2024).

Although the idea of object-centric gaze prediction is conceptually appealing and clearly relevant for downstream applications, progress in this direction has been hampered by data limitations. The GOO dataset is extremely constrained: only 5% of its images are real, which forces models to bridge a significant simulation-to-real (sim2real) domain gap (Xie et al., 2025), and all the scenes are captured in a retail setting, with people looking at products on shelves. As a result, methods trained on GOO fail to generalize to other domains and consistently underperform approaches trained on more traditional gaze following benchmarks. Refer to Figure 2.6 for a sample image and its corresponding annotations.

Beyond dataset bias, the task itself poses intrinsic challenges. In cluttered retail environments, the level of aleatoric uncertainty (i.e. irreducible error; Kendall and Gal (2017)) is high: often the visual evidence is insufficient to disambiguate exactly which product on a shelf a person is fixating on, even when the eyes are visible. In such cases, even a perfect algorithm trained on unlimited data would be unable to assign a unique bounding box, since the gaze cone would naturally cover multiple stacked or adjacent items. More broadly, this difficulty is not limited to retail scenes: many real-world environments involve dense arrangements of similar or overlapping objects (e.g. books on a desk, tools in a workshop), where gaze cues are simply too coarse for object-level resolution.

Another fundamental limitation is that object bounding boxes are only well-defined for things (countable objects such as a book or a cup) and not for stuff (uncountable regions such as the ocean, the sky, or a wall). In these latter cases, bounding-box formulations are not meaningful, whereas a gaze point remains universally applicable: it can represent both discrete objects and amorphous regions with equal validity. Even for things, the granularity of the representation depends on object scale. For small objects, the bounding box effectively collapses to a single point, making the gaze point alone sufficient. For larger objects, however, bounding boxes can complement point annotations; for example, a table may be annotated as a single object, but the gaze point disambiguates which side or which specific area of the table a person is looking at. This illustrates why gaze points remain essential, even in object-centric formulations.

Finally, annotating gaze targets at the object level is itself a major bottleneck. Unlike point annotations, bounding boxes for gaze targets require fine-grained judgments about object boundaries, which can be ambiguous (e.g. is a person looking at a whole shelf or a single bottle?). This annotation overhead discourages large-scale dataset creation outside of narrow domains.

In summary, while object-centric gaze prediction remains a promising research direction, its progress has been constrained by dataset limitations, annotation challenges, and the inherent uncertainty of the task. Without broader and more diverse datasets, methods in this area are unlikely to generalize beyond specialized contexts.

2.3.3 Social Gaze Prediction

Definition & Scope. Social gaze signals (i.e. eye contact, shared attention, and gaze patterns in dyads) are core building blocks of human communication. Developmental and cognitive science highlight their special status: infants follow others' gaze by the end of the first year and show early preferences for direct gaze; across species, social gaze supports coordination, learning, and affiliation (Scaife and Bruner, 1975; Farroni et al., 2002; Emery, 2000; Kleinke, 1986; Mundy and Newell, 2007). These insights motivate computer vision formulations that aim to infer when two people look at each other, when groups converge on a common target, or when someone looks back at the camera/robot. Each of these is useful for social scene understanding, HRI, and autonomous driving.

Mutual Gaze. Early work formalized "Looking At Each Other" (LAEO) in videos as detecting whether and when a pair of tracked heads engage in mutual gaze, coupling head pose dynamics with relative geometry (Marin-Jimenez et al., 2014). Subsequent CNN models reason over short tracklets and relative position maps (LAEO-Net), and later over longer temporal windows with improved tracking and training (LAEO-Net++), with benchmarks such as UCO-LAEO and AVA-LAEO enabling reproducible evaluation (Marin-Jimenez et al., 2019, 2022). Beyond two-stage pipelines that rely on given head tracks, Doosti et al. (2021) showed that auxiliary (pseudo) 3D gaze cues can boost LAEO classification, while Guo et al. (2022) recasts LAEO

as a one-stage set prediction task, detecting head boxes and mutual-gaze pairs jointly via a DETR-like transformer.

Shared Attention & Joint Attention. Shared attention seeks to infer whether multiple people simultaneously attend to the same target and, in some settings, to localize that target. Fan et al. (2018) introduced the VideoCoAtt dataset and a spatio-temporal network that fuses per-person gaze cones with region proposals to detect shared-attention intervals and predict target locations; a follow-up graph-based approach explicitly reasons about "gaze communication" structure over time (Fan et al., 2019). Later works explored end-to-end formulations and contextual reasoning: Attention Flow learns joint attention directly from RGB with saliency-augmented attention mechanisms (Sumer et al., 2020), while Nakatani et al. (2023) model single-frame joint attention with a transformer that integrates people attributes (locations, actions) as contextual cues to produce pixel-wise confidence maps. Recent efforts also consider temporal transformers for recognizing frame-level "atomic" behaviors (e.g. Mutual, Share) and aggregating them into event-level communication (Fan et al., 2019; de Belen et al., 2023).

Egocentric Eye Contact. A related line of work detects whether a person is looking toward the camera/agent, which is important for HRI and for anticipating pedestrian intent. Kotseruba et al. (2016) released JAAD (Joint Attention in Autonomous Driving): an in-the-wild driving clips with social/behavioral annotations and has been used to study driver–pedestrian communication. Building on this, Belkada et al. (2021) introduced LOOK, a cross-dataset benchmark for eye-contact detection in the wild (annotating pedestrians across multiple autonomous-driving datasets) and showed that pose/keypoints generalize better than raw-image models given the low resolution images and the distance of pedestrians from the camera. Complementary work proposes geometric heuristics for estimating driver–pedestrian eye contact from relative eye vectors (Onkhar et al., 2021) and alternative video-mining strategies for eye-contact segmentation "in the wild" (Wu and Sugano, 2022).

From a developmental-psychology standpoint, Chong et al. (2020a) trained a high-accuracy eye-contact classifier on more than four million annotated frames from videos of child–caregiver interactions, although the dataset itself was not released. More recently, Hempel et al. (2024) introduced NITEC: the largest in-the-wild eye-contact image dataset to date. By holding the architecture fixed, the authors reported higher accuracy than Chong et al. (2020a).

Unified Social Gaze Formulations. Beyond task-specific models, several datasets/frameworks aim to unify social gaze subtasks under common labels and protocols. The VACATION dataset defines "atomic-level" gaze communication states (e.g. Single, Mutual, Avert, Refer, Follow, Share) and "event-level" composites, paired with a spatio-temporal graph model to infer these structures (Fan et al., 2019). However, their annotation protocol is limited to a single gaze state per subject, even though two can occur at the same time (e.g. mutual gaze between A and B, and shared attention between A and C toward B; *cf.* Figure 2.7). More recently, de Belen et al. (2023) proposed a transformer-based architecture equipped with a temporal module to infer the same atomic/event labels end-to-end, without requiring head/object proposals.



Figure 2.7: Images from the VACATION dataset where two gaze states co-occur for the green subject, i.e. shared attention (SA) with the blue subject and mutual gaze (MG) with the pink one. However, only one state can be assigned to the green subject according to Fan et al. (2019)'s annotation scheme.

In the dyadic regime, Chang et al. (2023) first pointed out that VACATION annotation labels had overlapping definitions, and proposed the GP-Static dataset by re-annotating videos from VACATION and UCO-LAEO with five static mutually exclusive gaze patterns, providing a clean testbed for dyadic social gaze beyond LAEO alone. Together, these resources help connect mutual gaze, shared attention, and camera-directed eye contact under a single semantic umbrella, facilitating cross-task learning and evaluation.

Discussion and trends. Across tasks, three modeling axes recur. First, *pairwise vs. global*: LAEO methods often operate on person pairs or tracks; shared attention requires multi-person reasoning to partition people into one or more jointly-attending sets. Second, *explicit geometry vs. learned context*: classic pipelines inject gaze cones, depth, and pairwise geometry; newer transformers favor end-to-end learning over scene tokens and people tokens, sometimes with auxiliary cues (pose, actions) to disambiguate intent. Third, *one-stage vs. two-stage*: recent one-stage "instance detection" formulations (e.g., MGTR for LAEO) offer simplicity, efficiency and set-based matching, whereas two-stage approaches offer modularity and benefit from accurate detectors/trackers and straightforward supervision. Meanwhile, the psychology literature continues to motivate richer labels (mutual, referential, following) and temporal reasoning, encouraging unified datasets and metrics that better reflect social semantics rather than only pixel distances or frame-wise pair labels.

3 Datasets & Benchmarks

In the previous chapter, we reviewed the theoretical background and existing approaches to gaze understanding, with a particular focus on the tasks of gaze following and social gaze prediction. Building on this foundation, we now shift our focus to the datasets and benchmarks that underpin research in this area. The availability and quality of datasets have been central to the progress of the field: they determine not only what models can learn, but also how fairly and comprehensively we can evaluate them.

In this chapter, we present the datasets and benchmarks used throughout this thesis, which fall broadly into two categories: gaze target prediction (often framed as gaze following) and social gaze prediction, where the emphasis is on modeling interpersonal gaze behaviors. For each dataset, we discuss its motivation, data collection and curation strategy, annotation protocol, and key statistics, before situating it within the broader landscape of benchmarks.

Beyond surveying existing resources, this chapter also introduces the datasets that were developed as part of this thesis. These contributions were motivated by specific limitations in prior benchmarks and aim to enable a more complete and socially grounded study of gaze behavior. Together, these resources form the empirical foundation for the models and analyses developed in the remainder of this thesis. We summarize these datasets in Table 3.4.

3.1 Gaze Target Prediction

3.1.1 GazeFollow

Description. The GazeFollow dataset was introduced in Recasens et al. (2015) to study gaze target prediction in unconstrained settings. It aggregates over 120k images and 130k annotated gaze instances from diverse sources such as SUN (Xiao et al., 2010), MS-COCO (Lin et al., 2014), Pascal VOC (Everingham et al., 2010), ImageNet (Deng et al., 2009) and Places (Zhou et al., 2016), making it the first large-scale effort to capture where people look in natural images. It features a high diversity of indoor/outdoor scenes, people, objects, and activities. This



Figure 3.1: Samples from the GazeFollow dataset. The images feature diverse people and settings, different aspect ratios, and a relatively small resolution. The rightmost image is from the test set, which explains why it has 10 gaze point annotations from different labelers.

explains why the dataset has since become the de facto benchmark for gaze following. We provide a few samples in Figure 3.1 for reference.

Annotations. The dataset was annotated via Amazon Mechanical Turk (MTurk), where a gaze annotation instance is a person's head bounding box, the center point of their eyes, and a gaze point that localizes where they are looking. The test set, which includes 5k instances, was annotated with ten gaze points, while keeping the head box fixed, to quantify inter-annotator consistency and account for the inherent ambiguity of gaze. The annotations were later extended by Chong et al. (2020b) with an inside/outside binary class. Most instances are looking inside the image which explains why models do not include the *in/out* prediction when training on GazeFollow.

Limitations. The dataset images are of relatively low resolution, typically around 400–600 pixels per side, and many suffer from compression artifacts that reduce visual clarity. Moreover, in multi-person scenes, human labelers tend to favor individuals in the foreground, where gaze is easier to resolve, which introduces a bias against those positioned farther from the camera. Finally, the annotations also lack the precision of more recent datasets; a limitation that can be partly attributed to the use of the crowdsourcing platform Amazon Mechanical Turk, which is known to yield noisier labels compared to expert annotation.

Semantic Extension. A core contribution of this thesis is the introduction of semantic grounding, moving beyond pixel coordinates to also predict the class of the gaze target; a topic we will explore in detail in Chapter 8. To enable models to jointly predict a gaze location and its semantic class, we extended the GazeFollow dataset (Tafasca et al., 2024a) with class labels for each gaze instance. This required designing a pseudo-annotation pipeline that automatically infers the semantic label of the ground-truth gaze target. Since gaze following datasets provide only point annotations, our strategy was to first segment each image and then associate the gaze point with the semantic class of the pixel it falls on (see Figure 3.2).

Because full-scene segmentation is necessary, it was crucial to rely on models with open-vocabulary capabilities able to describe arbitrary objects. To this end, we combined two recent open-source segmentation systems built on foundation vision–language models. The first,

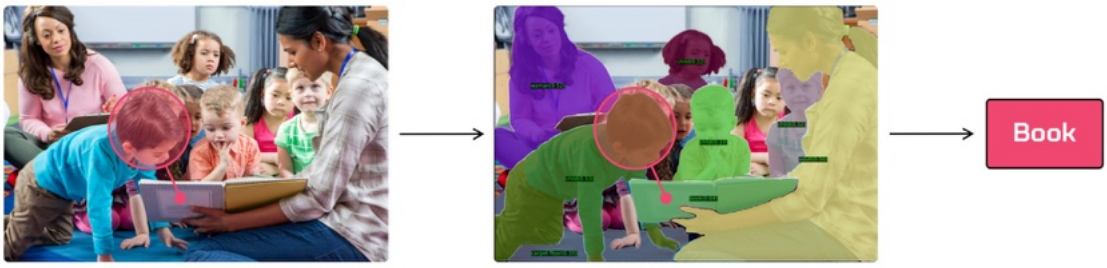


Figure 3.2: Semantic pseudo-annotation pipeline. The idea is to match the ground-truth gaze point with the pseudo-label of the underlying pixel produced by an open-world segmentation system.

RAM-Grounded-SAM (Kirillov et al., 2023; Liu et al., 2023; Ren et al., 2024), produces highly precise masks and reliable labels, but often misses regions altogether. To compensate, we employed Semantic-Segment-Anything (Chen et al., 2023), which provides broader coverage but at the cost of noisier labels and frequent over-segmentation, where one object is fragmented into multiple small parts. A side-by-side comparison of the two methods is shown in Figure 3.3.



Figure 3.3: Comparison of the two open-world segmentation methods used to derive the pseudo-labels for GazeFollow: RAM-Grounded SAM (RGS) and Semantic-Segment-Anything (SSA).

Applying this pipeline to GazeFollow yielded roughly 5000 unique raw labels. We then performed extensive text processing: correcting incomplete words, removing stopwords, duplicates, adjectives, and prepositions, which reduced the vocabulary to about 3700 cleaned pseudo-labels.

To establish a reliable benchmark for evaluation, we further manually annotated the test set with gaze target classes. Since semantic gaze annotation cannot rely on a closed vocabulary established in advance—because it is impossible to predefine which objects people will be looking at without first annotating their gaze—we adopted an open-set approach. Within this open setting, we still enforced internal consistency (e.g. avoiding synonyms whenever possible) by using a single expert annotator for the entire test set; an idea that was used before for datasets like ADE20K (Zhou et al., 2017). Moreover, since gaze heatmaps often overlap with multiple objects, we adopted a multi-label annotation strategy: for instance, a person cutting a cake may simultaneously be looking at the cake and the knife. This approach also helped address hierarchical ambiguities, such as considering both “wheel” and “car” as valid targets when the gaze falls on the wheel. The resulting test set contains a curated vocabulary of 346 classes, with Figure 3.4 presenting a representative word cloud. We also provide some samples illustrating both train set pseudo-labels and test set manual ground-truth labels in Figure 3.5.



Figure 3.4: Word clouds of the vocabulary of gaze labels. **Left:** pseudo-labels of GazeFollow. **Middle:** ground-truth labels of the test set of GazeFollow. **Right:** ground-truth labels of GazeHOI.



Figure 3.5: Samples from the GazeFollow dataset with semantic annotation (i.e. training set pseudo-labels and test set ground-truth labels). For test set samples, we only show a single gaze point and omit the remaining ones.

The pseudo-labels, while noisy, are very diverse.

3.1.2 VideoAttentionTarget

Description. The VideoAttentionTarget dataset was introduced by Chong et al. (2020b) as a video-based complement to GazeFollow. It is the first large-scale dataset for gaze following in videos, comprising 1,331 clips extracted from 50 popular TV shows and YouTube series, for a total of 72k high-resolution frames (typically 1920×1080). These clips capture a wide range of human interactions, from casual conversations to more dynamic exchanges. An example is shown in Figure 3.6.

Annotations. As in GazeFollow, each annotated instance consists of a subject head bounding

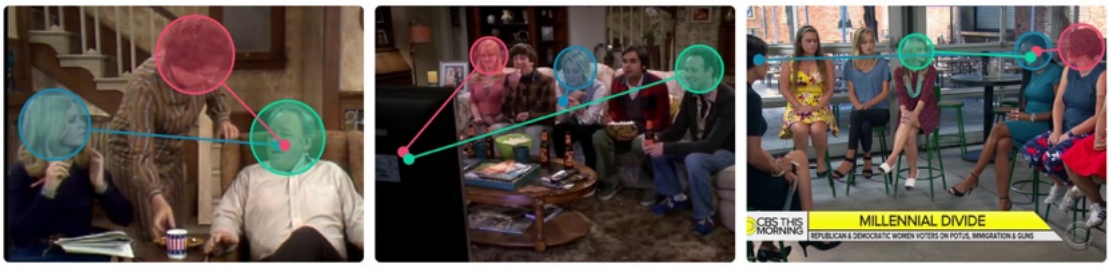


Figure 3.6: Samples from the VideoAttentionTarget dataset. The frames are high-resolution and annotations are sufficiently accurate. The number of shows/clips is small, which limits the diversity of scenes and people. We show 3 annotated instances per frame, each with a head box and gaze point. Subjects looking outside do not have a gaze point.

box, a gaze point, and an *in/out* label indicating whether the target lies inside or outside the frame. A key difference, however, is that VideoAttentionTarget provides annotations for most visible people in each frame, resulting in a richer set of 164k annotated instances across the dataset.

Limitations. While both the image quality and annotation accuracy are generally higher than in GazeFollow, the dataset suffers from lower scene diversity. Despite the larger number of annotated instances, the content is constrained by its video-based nature and its focus on TV shows. As a result, the majority of gaze targets fall on other people's faces rather than on objects (*cf.* Table 3.2). Moreover, the cinematographic style of TV productions introduces additional biases: narrow fields of view, recurring camera angles, and limited background context. These distributional constraints limit the ecological diversity of the data, which is essential for training gaze prediction models that generalize well.

3.1.3 DLGaze

Description. The DLGaze dataset was introduced by Lian et al. (2018) as a small-scale video-based gaze following dataset. It records 16 volunteers performing daily life activities (e.g. talk, read, use a mobile phone) in 4 different indoor environments (i.e. laboratory, working office, library, and corridor). The dataset consists of more than 5k frames annotated with close to 10k gaze instances. Some samples from the dataset are shown in Figure 3.7.

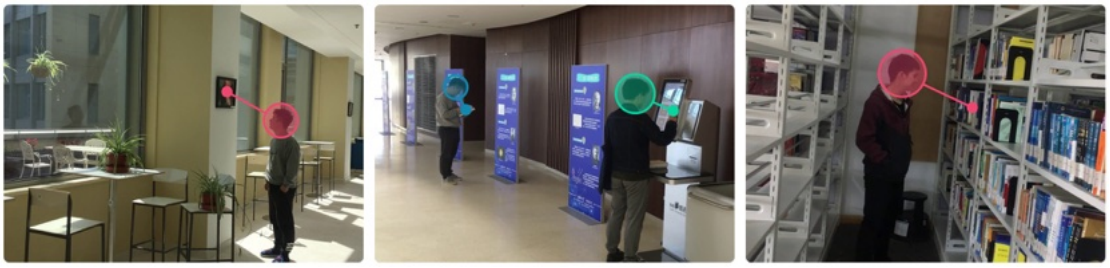


Figure 3.7: Samples from the DLGaze dataset. The images depict participants performing daily activities in different indoor environments. This is very different from GazeFollow, which makes it suitable for cross-dataset evaluation and the assessment of model generalization.

Annotations. The dataset provides for each frame the head bounding box and the gaze point coordinates. Unlike previous datasets, the gaze point is derived from observers (i.e. people in the video) rather than inferred (i.e. guessed) by external annotators.

Limitations. While DLGaze improves realism and reliability through observer-annotated gaze, it remains constrained by typical video-based limitations: the dataset is considerably smaller and offers less diversity in scenes and participants than VideoAttentionTarget. Consequently, and because its visual domain differs substantially from previously introduced datasets, it is often used not for training but as a cross-dataset benchmark to evaluate how well models trained on larger datasets generalize.

3.1.4 ChildPlay

Motivation. Understanding gaze behavior in children is central to developmental psychology and clinical research, as it underpins social communication skills such as joint attention and eye contact, and is a key marker for disorders like autism. For this reason, several methods have been proposed to analyze children's gaze and behavior, particularly in the context of autism (Shic et al., 2011; Rehg et al., 2013; Li et al., 2016; Chong et al., 2017; Syeda et al., 2017; Hashemi et al., 2018). However, most existing public gaze benchmarks have been collected in restricted environments, and almost exclusively feature adults. This bias matters: first, there is a marked difference in gaze behavior between adults and children (Franchak et al., 2016), and second, it is well-documented in adjacent tasks like human pose estimation that models trained on adult data underperform on children (Sciortino et al., 2017), reflecting a real distribution shift (body scale/proportions, head–torso geometry, typical scenes) that also impacts gaze, where head pose and morphology are key cues (Guan et al., 2019; Belkada et al., 2021; Gupta et al., 2022). At the same time, datasets that do focus on children are typically private due to privacy concerns (De Belen et al., 2020), restricting reproducibility and slowing down progress in the field.

To address this gap, we introduce ChildPlay (Tafasca et al., 2023b), the first publicly available benchmark dedicated to predicting the gaze of children in naturalistic interactions with adults. The dataset consists of short video clips recorded in childcare facilities, schools, and therapy centers, capturing unscripted play and learning activities in rich, unconstrained environments. By releasing this dataset, we aim to provide the community with a resource that enables the study of children's gaze behavior in-the-wild and fosters the development of models that generalize beyond adult-dominated benchmarks to child-specific applications. We provide a few samples to showcase the features of ChildPlay in Figure 3.8.

Data Collection Strategy. We collected source material from YouTube using queries such as "children playing toys", "childcare center", and "kids observation" to retrieve videos relevant to our aims. To ensure visual quality, we only considered videos with a 16:9 aspect ratio and a resolution of 720p or 1080p. Audio tracks were downloaded as well, though in many cases they



Figure 3.8: Samples from the ChildPlay dataset. The frames are high-resolution and high-quality. The number of original clips is higher than VideoAttentionTarget but remains limited in general. We show all annotated instances per frame, each with a head box and gaze point. Subjects looking outside do not have a gaze point.

consisted of commentary or unrelated background sound rather than scene audio.

The collected videos span diverse contexts, including childcare facilities, schools, homes, and therapy centers. Because raw videos often contain long irrelevant segments, we extracted shorter clips that feature children and exhibit clear gaze shifts or social interactions. We excluded clips with scene cuts, strong blur, heavy overlays, excessive zoom, or unstable camera movement. To encourage variety, we also limited both the number and duration of clips sourced from any single video or YouTube channel.

The clips in the dataset primarily depict indoor environments, and contain at least one child, often alongside one or two adults and additional children. The children range in age from toddlers to pre-teens. While the predominant activity is toy play, the dataset also includes other forms of interaction, such as structured behavioral therapy exercises.

Annotation Protocol. A major limitation of prior gaze-following datasets is their coarse target taxonomy, typically restricted to a binary distinction between *inside* and *outside* the frame. Real-world scenes, however, present richer edge cases that this dichotomy fails to capture. Examples include *eyes closed* (no valid target), *occluded targets* (in-frame but not visible), and *gaze shifts* (transitions without a stable fixation). We illustrate these cases in Figure 3.9. ChildPlay explicitly incorporates these cases into its annotation protocol, aligning the dataset more closely with how gaze unfolds in the wild and enabling additional tasks such as gaze shift detection.

We conducted dense frame-level annotations. In each clip, up to three people were selected and, for every frame, annotated with a head bounding box, a 2D gaze point, a gaze label, and a binary child/adult tag. Crucially, the gaze label extends beyond the standard inside/outside distinction. In earlier datasets such as VideoAttentionTarget, intermediate frames during a gaze shift were often labeled as *outside*, a choice inconsistent with the intended definition. To avoid such ambiguities, we defined a taxonomy of seven mutually exclusive classes that account for these special scenarios (*cf.* Table 3.1).

In practice, the dataset remains dominated by inside-frame (85.3%) and outside-frame (5.4%) annotations, but the additional classes capture 9.3% of the data where a confident 2D gaze

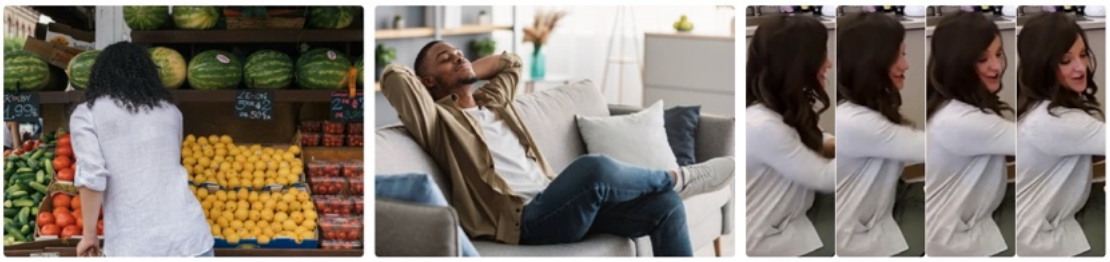


Figure 3.9: Example scenarios encountered in natural, in-the-wild images that are not adequately captured by the traditional inside vs. outside labels. **Occluded targets:** the gaze lies within the frame but is blocked by the subject's own body. While the approximate target can be inferred, an explicit flag is needed to properly handle such cases. **Closed eyes:** neither the inside nor outside categories apply, requiring a dedicated class to account for blinks or prolonged eye closure. **Gaze shifts:** during a transition from one target to another, the subject's head and eyes are in motion but not fixated on any meaningful intermediate point. Such moments should not be annotated with a continuous interpolated gaze location.

point could not otherwise be assigned.

To assess annotation reliability, we measured inter-annotator agreement on 2,000 double-coded instances. Following standard practice, we treated one annotator as the ground truth and evaluated the other as a prediction. The resulting human baseline is a distance error of 0.051 for target localization and an average precision of 0.925 for the in/out classification—values consistent with the agreement levels reported for VideoAttentionTarget.

Gaze Class	Definition
Inside Frame	The gaze target is located within the frame and is visible.
Outside Frame	The gaze target is outside the frame's boundary.
Gaze Shift	The person shifts their attention from one location to another over at least two consecutive frames. If needed, shorter shifts such as saccades can be recovered by detecting abrupt changes in gaze points within those annotated as `inside frame`.
Occluded	The 2D gaze target is within the frame but is totally occluded.
Uncertain	The gaze target cannot be determined confidently (i.e. lack of salient elements in the gaze direction, several possible targets, etc.).
Eyes Closed	Used in the rare cases where a child closes their eyes (e.g. during hide-and-seek).
Not Annotated	None of the other options is applicable.

Table 3.1: Definitions of the seven non-overlapping gaze classes used to annotate ChildPlay.

Statistics. The dataset consists of 401 short clips (average length of 10 seconds), extracted from 95 videos originating from 44 distinct YouTube channels. In total, it contains 258k annotated instances (62% involving children), distributed over 120k frames. Annotations were produced by seven annotators using the LabelBox platform (Labelbox, 2025).

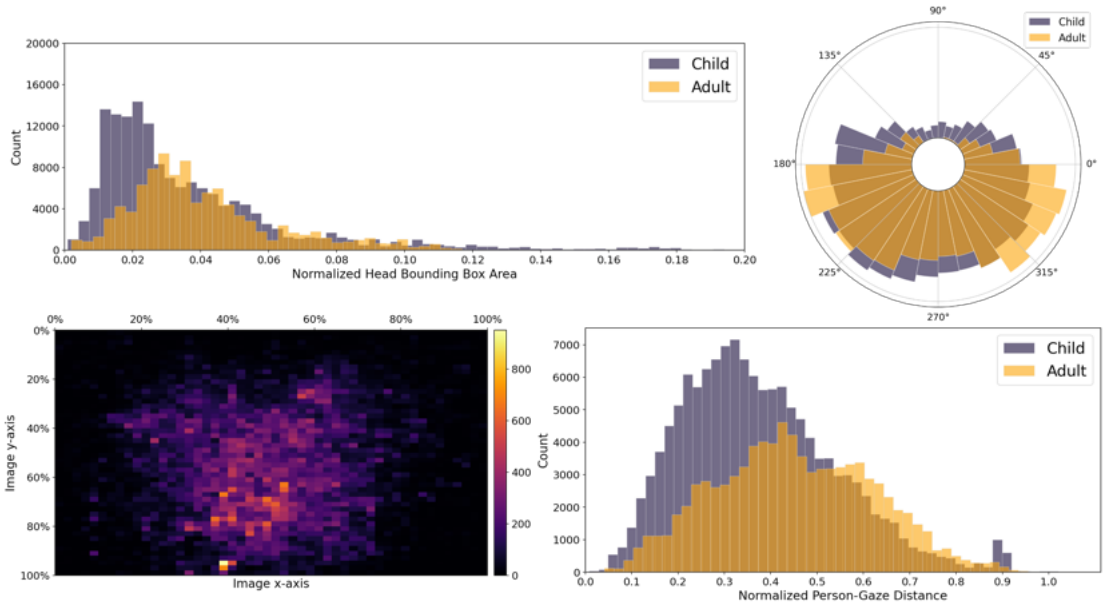


Figure 3.10: Geometric statistics of the ChildPlay annotations. **Top-Left:** Distribution of head bounding box area normalized w.r.t image size. **Top-Right:** Distribution of gaze angles in the image frame. **Bottom-Left:** 2D histogram of gaze points. **Bottom-Right:** Distribution of the distance between each person (i.e. center of the head) and their gaze point, normalized w.r.t the image sides.

Figure 3.10 summarizes the distribution of key geometric quantities, revealing clear differences between children and adults. Head sizes span a wide range, reflecting the varied distances of participants from the camera. A consistent behavioral pattern also emerges: adults often look downward to monitor children and their activities, while children tend to focus on nearby toys but occasionally lift their gaze to engage with adults. This distinction is further supported by the large disparity in the probability of looking at faces (*cf.* Table 3.2). Finally, the distance between head and gaze point is usually between 10% and 60% of the image side, with children concentrating more on close targets compared to adults.

% LAH	GazeFollow	VideoAttentionTarget	ChildPlay (C)	ChildPlay (A)
All	23.0	69.0	15.7	44.4
Multi-person	30.6	71.0	16.9	44.6

Table 3.2: Percentage of Looking-At-Heads (LAH) instances across datasets. An LAH instance is where the annotated subject is looking at another person's head. For ChildPlay, **C** refers to the children subset and **A** to the adults subset.

Dataset Comparison. When compared to VideoAttentionTarget (VAT), ChildPlay provides complementary content. VAT primarily features adults in TV-style recordings, where people are often standing, sitting, or conversing, and gaze targets are strongly biased toward other people's faces. In contrast, ChildPlay emphasizes children and captures a broader range of poses, activities, and social dynamics. For example, children looking down at toys, or adults looking down at children. This results in more diverse gaze behaviors and richer contextual cues. A semantic analysis of gaze targets (*cf.* Table 3.3) confirms this distinction: ChildPlay is

far more balanced[1] across categories, while also offering 50% more frames and nearly twice the scene variety of VAT.

	Things (Person)	Things (Other)	Stuff	Not Detected
VideoAttentionTarget	80.85%	80.05%	3.60%	7.50%
ChildPlay	45.19%	18.66%	12.62%	23.53%

Table 3.3: Comparison of the gaze target semantic class across datasets. Numbers were obtained by running the panoptic segmentation model from Cheng et al. (2022) on images and retrieving the semantic class of each annotated gaze point.

Relative to GazeFollow, ChildPlay trades off some breadth for quality and realism. GazeFollow remains unparalleled in its scale and diversity of static images, but it suffers from lower resolution, compression artifacts, and noisier annotations. Moreover, image-based datasets are often biased toward people posing for the camera, which reduces ecological validity. By contrast, ChildPlay's video clips—sourced from unscripted, everyday settings rather than cinematographic productions—capture gaze in natural, dynamic, and often cluttered environments. This makes the dataset not only higher in visual and annotation quality, but also more representative of in-the-wild gaze behavior, providing unique challenges and opportunities for computer vision models.

Extensions. Since its release, ChildPlay has been well received by the research community, attracting attention from computer scientists, psychologists, and behavioral scientists alike. At the time of writing, it has been downloaded more than 400 times, reflecting strong interest in child-centered gaze benchmarks. Beyond its original annotations, several works have built on ChildPlay by adding complementary behavioral layers, thus expanding its utility across disciplines.

Recent efforts include the addition of *speech and vocalization labels* (Gupta et al., 2024b), which make it possible to study the coordination between gaze and spoken communication in children and adults. Other researchers have annotated blinks and head gestures, enabling analyses of multimodal cues that often accompany gaze behaviors. The ChildPlay-Hand extension (Farkhondeh et al., 2024) provides detailed *annotations of hand-object interactions*, linking gaze with manipulative actions and enriching opportunities for studying triadic attention (child–adult–object). Building on this, Penchala et al. (2025) introduced *engagement and collaboration annotations*, allowing models to move from gaze detection to higher-level behavioral inference. Finally, Boluk and Kose (2024) contributed *autism-related diagnostic labels*, highlighting the dataset's clinical relevance and its potential role in developing early screening tools.

Together, these extensions demonstrate the versatility of ChildPlay as a platform for multimodal behavioral research. By layering gaze with speech, gestures, interactions, and diagnostic

[1] A manual inspection revealed that most of the `not detected` instances in ChildPlay shown in Table 3.3 correspond to objects belonging to the `things (other)` category that were not detected by the segmentation.

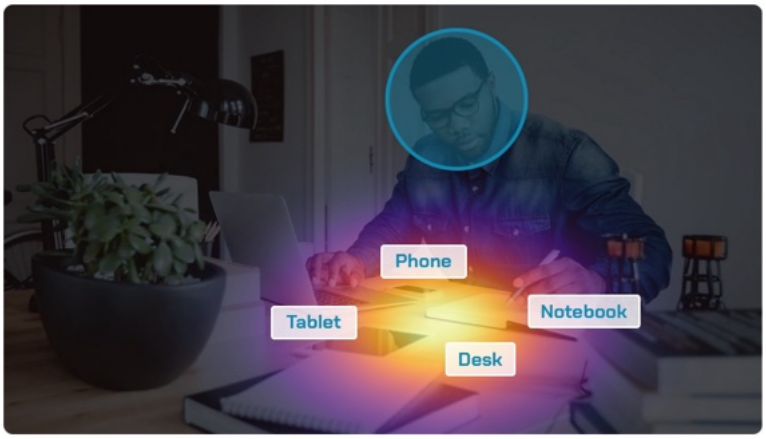


Figure 3.11: Example of the limitation of point-based annotation and evaluation. The predicted gaze heatmap overlaps multiple nearby objects (phone, notebook, tablet, desk). A distance-based metric cannot distinguish whether the model selects the correct semantic target, whereas adding class labels makes it possible to evaluate semantic correctness (e.g. looking at the phone vs. the notebook).

signals, the dataset now supports investigations ranging from fundamental developmental psychology to applied machine learning and clinical screening.

3.1.5 GazeHOI

Motivation. GazeHOI is the second dataset contribution of this thesis (Tafasca et al., 2024a). Its role is to provide a clean, human-verified testbed for evaluating *both* gaze target localization and the *semantics* of the target. This complements our GazeFollow semantic extension, where training relies on pseudo-labels. In practice, many applications care less about a raw (x, y) coordinates and more about *what* is being looked at. Yet, all popular gaze following benchmarks lack target classes, making it hard to assess whether a model selects the right *object* when several plausible locations are nearby. This ambiguity arises from an interplay between dataset and evaluation: point-only annotations force metrics to collapse semantic correctness into geometric proximity, which can mis-rank models in cluttered scenes. We illustrate this scenario in Figure 3.11.

Why hasn't the field simply added class labels before? Because target labeling presupposes target localization. If you annotate from scratch "in the wild," you cannot predeclare a closed vocabulary and guarantee people only look at those classes; i.e. you must allow open-set labels. That creates two practical problems: (i) enforcing consistent naming across annotators (e.g. creek vs river), and (ii) covering the long tail of uncommon or "stuff" categories that are hard to name reliably. You can solve this by brute force: first localize everything people look at and then constrain the vocabulary, or hire language experts to normalize an open vocabulary. However, both paths are expensive to pursue at scale, both financially and time-wise. These constraints are discussed in several prior works (Krishna et al., 2017; Gupta et al., 2019; Kuznetsova et al., 2020) and motivated our design choices here.

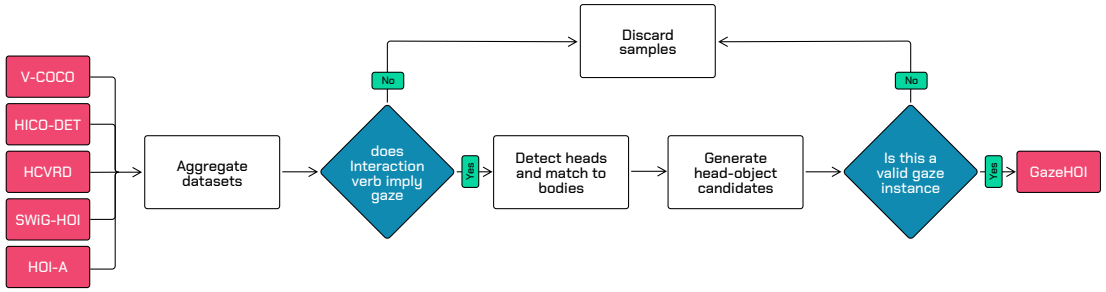


Figure 3.12: Workflow diagram of the GazeHOI construction process.

Data Construction. The central idea behind GazeHOI is to circumvent the cost and complexity of building a dataset from scratch by re-purposing existing Human–Object Interaction (HOI) corpora. Instead of confronting the open-vocabulary and consistency challenges that make semantic gaze annotation prohibitively expensive, we leverage mature HOI datasets that already provide reliable person and object bounding boxes together with object class labels. This choice is conceptually natural: human–object interaction datasets are designed to capture people engaging with their surroundings, which is typically accompanied by rich gaze behaviors. In many cases, looking at an object can itself be viewed as a form of interaction. In fact, some datasets, such as V-COCO (Gupta and Malik, 2015), even include an explicit interaction verb labeled `looking at`.

Concretely, we aggregate five large-scale HOI datasets: V-COCO (Gupta and Malik, 2015), HICO-DET (Chao et al., 2018), HCVRD (Zhuang et al., 2018), SWiG-HOI (Wang et al., 2021), and HOI-A (Liao et al., 2020). Next, we filter them to retain verbs likely to imply visual attention (e.g. *cut*, *repair*, but not *carry*), detect and match head regions to the annotated persons' body bounding boxes, and then apply a lightweight human verification stage to confirm whether the person is indeed looking at the labeled object. Figure 3.12 shows a workflow diagram of this process.

This approach is an example of *dataset bootstrapping*: a practice increasingly adopted in computer vision (Mohammed et al., 2023; Chen and Hong, 2024), where new benchmarks are built by extending or re-purposing existing corpora; e.g. RefCOCO (Mao et al., 2016) from COCO (Lin et al., 2014). By collapsing the heavy lifting of annotation into a rapid verification step, GazeHOI demonstrates how clever repurposing can yield large-scale, semantically grounded resources at minimal cost. The result is a diverse dataset with almost 45k images and 60k instances, some of which are shown in Figure 3.13.

Annotation Pipeline. A major factor enabling this efficiency was our custom-built, locally deployed annotation tool. For each candidate image, the tool overlaid the head bounding box and the object bounding box from the HOI annotation, allowing the annotator to quickly decide whether the gaze aligned with the object. Local deployment proved critical: it avoided the latency and bandwidth bottlenecks of cloud-based labeling platforms, where loading a high-resolution image can take seconds even if the verification itself requires less than one. By streaming images directly from local storage, the tool reduced annotation to its



Figure 3.13: Samples from the GazeHOI dataset. The images feature high diversity in scenes, people, objects, and interactions. The resolution and aspect ratios vary across images. We show both the bounding box and the class label of the object being interacted with, but we omit the interaction verb.

essence—decision speed—and allowed one of the annotators to process approximately 80k candidates in just 15 days.

In short, GazeHOI showcases how a scalable, diverse, and semantically grounded gaze benchmark can be constructed in under 20 days by combining dataset bootstrapping with a streamlined local verification pipeline, turning what is usually months of costly annotation into a fast, practical, and resource-efficient process.

Statistics. The final GazeHOI dataset contains 44k images and 58k gaze instances, after discarding roughly 50% of candidate images during the verification stage. Each instance is annotated with the person's body and head bounding boxes, the object bounding box, the object class, and the interaction verb. The dataset vocabulary spans 985 object classes, of which 522 rare categories (fewer than 10 instances each) are isolated into a separate split. This design allows the dataset to serve not only as a benchmark for standard gaze target prediction, but also as a testbed for future research on open-vocabulary semantic gaze following, in line with practices from open-vocabulary object detection (Gu et al., 2021). A word cloud of the object classes from GazeHOI was presented in Figure 3.4.

For the experiments reported in this thesis, we focus on a curated subset comprising the remaining 463 object classes and 42k images (i.e. 56k instances), balancing scale with reliability. Finally, to prevent overlap with existing gaze following resources, we applied a de-duplication pipeline that ensures no images in the validation or test sets of GazeHOI appear in the training splits of either GazeHOI or GazeFollow, given that both of them draw from popular vision benchmarks.

Dataset Comparison. To date, the only gaze following dataset with semantic target annotations is GOO (Tomas et al., 2021). However, as discussed earlier, GOO is 95% synthetic and limited to a simplified retail environment (i.e. the authors built a mock-up of a retail setting). Its scenes are dominated by 80 densely stacked shelf products, where object-level gaze resolution is often infeasible due to minimal visual and contextual cues. By contrast, GazeHOI provides over 43k diverse, real-world images in which people naturally interact with a wide variety of objects. This diversity makes GazeHOI not only a stronger benchmark for evaluating semantically extended gaze following methods, but also a valuable resource for

training models with practical downstream applications.

3.2 Social Gaze Prediction

3.2.1 UCO-LAEO

Description. UCO-LAEO is one of two datasets introduced in (Marin-Jimenez et al., 2019, 2022) for the task of detecting when two people are Looking At Each Other (LAEO) in videos. The dataset is drawn from four popular TV shows (Game of Thrones, Mr Robot, Silicon Valley, and The Walking Dead). From these, the authors extracted 129 shots (video clips) each about 3–12 seconds long, subdivided into more than 18k frames for a total of 36k mutual gaze instances. A set of samples from this dataset is given in Figure 3.14.



Figure 3.14: Samples from the UCO-LAEO dataset. The images are sourced from TV shows so they are similar to VideoAttentionTarget. We show positive mutual gaze instances in green and negative ones in pink.

Annotations. For UCO-LAEO, all heads visible in each frame are annotated with bounding boxes. Then, for each pair of heads in each frame, a binary label indicates whether the pair is in a mutual gaze (LAEO) or not. There is also an `ambiguous` label for uncertain cases, and a distinction in positive labels when visibility or clarity is compromised (for example, when one person's face is occluded).

Limitations. Although UCO-LAEO provides useful annotated data for studying mutual gaze in videos, it has several constraints. First, it is video-based with a modest overall size: only 129 shots from 4 TV shows limits the variety of environments, backgrounds, camera angles, lighting, and kinds of interactions. This makes generalization harder to other domains. Second, because the data comes from TV shows, there is bias toward staged, dialogic, and interpersonal settings rather than everyday, spontaneous mutual gaze. Finally, as explained before, the cinematic style tends to use narrower fields of view, relatively consistent lighting, and well-framed shots, which reduce visual clutter and variation.

3.2.2 VideoCoAtt

Description. VideoCoAtt is a dataset collected for studying shared attention in third-person video settings (Fan et al., 2018). The dataset contains 380 video sequences drawn from 20



Figure 3.15: Samples from the VideoCoAtt dataset. The clips are sourced from TV shows but they are distributed in low resolution and are heavily compressed so image clarity is average. We show positive shared attention instances, with the bounding boxes for subjects' heads as circles, and for the shared attention object as a rectangle. Different co-occurring instances are visualized in different colors.

different TV shows/movies spanning diverse cultures (American, Chinese, Indian, European, etc.). The video sequences range from ~20 seconds up to over a minute, at spatial resolution of 320×480. The frames include many social-scene settings: living rooms, kitchens, restaurants, offices, outdoors, etc., with varying numbers of people and instances of shared attention per frame. In total there are about 492k frames, and 140k positive shared attention instances. To obtain negative instances, we also detect additional heads in the scene that are not involved in shared attention. Any pair that includes at least one of these individuals is then automatically labeled as negative. We show a few samples from this dataset in Figure 3.15.

Annotations. Every frame is manually annotated using the online tool VATIC (Vondrick et al., 2013). For each frame, annotators first mark whether there is shared attention occurring in that scene. If yes, then they draw bounding boxes around each shared attention region (i.e. the object or human body part at which the group is looking) and also draw bounding boxes around all heads of people engaged in that instance of shared attention. This is repeated for all shared attention instances in a given frame. Non-visible or occluded shared attention targets are not annotated.

Limitations. While VideoCoAtt is large and richly annotated for shared attention, it has several limitations. First, the resolution is modest (320×480) and the frames are noticeably compressed, which limits detail (especially fine features like small objects or subtle gaze cues). Second, because the dataset is drawn from TV show content, many settings are staged or tend toward conversational scenes, which heavily biases the kinds of shared attention observed (e.g. people often look at each other or faces rather than at objects). Third, shared attention is only annotated when visible; occluded or out-of-view shared attention targets are excluded, which may miss realistic cases. Finally, despite the size in frames, the variety (lighting, camera angle, object diversity, cultural settings) is narrower than what one would expect for truly in-the-wild data, restricting how well models trained on it might generalize to very different domains.

3.2.3 VACATION

Description. VACATION (Video gAze CommunicATION) was introduced by Fan et al. (2019)



Figure 3.16: Samples from the VACATION dataset. The clips are sourced from the same TV shows as VideoCoAtt so the quality is relatively the same. We show both atomic-level (**A**) and event-level (**E**) gaze labels. Event-level labels depend on time, so the frames shown might might not be sufficient to resolve the behavior. Gaze target object bounding boxes are given in the same color. Arrows and small circles denote gaze direction and target centers. Lines between people's head boxes denote gaze relations. It is clear that the annotation scheme adopted by the authors is confusing and relies on overlapping gaze label definitions.

to study gaze communication in social videos at two levels: atomic (per-frame roles such as `single`, `mutual`, `avert`, `refer`, `follow`, `share`) and event (temporal compositions such as `non-communicative`, `mutual gaze`, `gaze aversion`, `gaze following`, `joint attention`). The corpus consists of 300 clips curated from more than 50 TV shows/movies, downsampled to 640×360. In total, there are 97k frames and 165k annotated instances.

Annotations. Frames are annotated with head bounding boxes and key objects' bounding boxes, where each person is linked to a single target (face or object). At the atomic level, each person receives one of the six labels per frame; at the event level, sequences are labeled with one of five communication events defined as temporal compositions of atomic states. Annotations were cross-validated by two annotators with adjudication, and the official train/test split avoids source overlap.

Limitations. Despite its rich structure, VACATION inherits several constraints. As explained in Chapter 2, the single-label-per-person design (explicit in the formulation) forces annotators to choose one atomic state even when multiple behaviors can co-occur (e.g. establishing mutual gaze while referring to an object), and the atomic definitions themselves overlap conceptually (e.g. `share` vs. `mutual` vs. `follow` within the same group) as argued in Chang et al. (2023). Specifically, the overlap arises mainly in two ways. First, the scheme mixes *relational states* with *temporal dependencies*: for example, after a follower shifts to the leader's target, the frame is simultaneously `follow` (temporal) and `share` (relational). Second, it conflates *target identity* with *target type*: the label `mutual` is defined as looking at each other's faces, while `share` is looking at the same target; yet if two people both look at a third person's face, the frame fits both. In this sense, the taxonomy slices behavior along non-orthogonal axes, forcing annotators to choose only one label even when several apply. Finally, the TV/movie source domain and 640×360 resolution narrow scene diversity and visual detail, potentially limiting transfer to in-the-wild settings. We provide samples from the dataset in Figure 3.16 to show image diversity, exemplify annotations, and illustrate some of the issues discussed before.

Name	Type	Shows	Frames	Instances	Origin	Annotations
GazeFollow	image	—	122K	130K	SUN, MS-COCO, ImageNet, ...	Eye location · 2D gaze point · inside/outside
VAT	video	50 (606)	72K	164K	TV shows	Gaze point · inside/outside
DLGaze	video	4 (86)	5K	10K	Manual collection	Gaze point
GOO	image	—	201K	172K	Manual collection & Synthetic	Gaze point · gaze object bbox & class
ChildPlay	video	95 (401)	120K	258K	YouTube	Gaze point · gaze class
GazeHOI	image	—	44K	56K	V-COCO, HICO-DET, HCVRD, ...	Gaze object bbox & class · interaction verb
UCO-LAEO	video	4 (129)	18K	36K	TV shows	Mutual gaze label
VideoCoAtt	video	20 (400)	492K	138K	TV shows	Shared gaze labels · object bbox
VACATION	video	50 (301)	97K	164K	TV Shows	Gaze object bbox · gaze communication label

Table 3.4: Summary of gaze estimation datasets. All datasets include annotated head bounding boxes. The *Shows* column indicates the number of original shows or domains each dataset draws from, with the total number of individual videos given in parentheses.

4 Experimental Protocol

Following the presentation of datasets and benchmarks in the previous chapter, we now turn our attention to the experimental protocol that governs how models are trained, evaluated, and compared. Establishing a rigorous and transparent protocol is essential to ensure fair evaluation across diverse datasets, to enable reproducibility, and to highlight the strengths and limitations of different approaches. In gaze following, in particular, protocol choices such as how ground-truth heatmaps are generated, how predictions are decoded into coordinates, or how context people are treated can substantially influence results. These details must therefore be carefully specified rather than left implicit.

In this chapter, we describe the metrics used to evaluate performance, the implementation details common to all methods, and dataset-specific training and evaluation routines for gaze following-related tasks. We also discuss adaptations required for datasets with unique annotation formats, such as gaze object bounding boxes, and motivate the use of pre-training on large-scale datasets like GazeFollow to address the limited diversity of video-based benchmarks. By systematically outlining these elements, this chapter sets the foundation for the experimental results presented later and ensures that our findings can be interpreted and reproduced with clarity. Unless stated otherwise, the methods described in subsequent chapters follow this default protocol, and any deviation from it will be explicitly highlighted. As for Chapter 7 on social gaze behavior, it will instead include its own dedicated protocol to avoid overloading the present chapter with task-specific details that are only relevant in one place.

4.1 Performance Metrics

In this section, we present and discuss the various metrics used to evaluate gaze target prediction methods, and summarize them in Figure 4.1.



Figure 4.1: Distance: measures single-target localization error as the normalized L2 distance d between the predicted ($\bar{y}$) and ground-truth (y) gaze points. The pink ellipse illustrates the set of predictions yielding the same distance d, highlighting the distortion introduced by the metric when the image aspect ratio is not square. **AUC:** evaluates performance in multi-modal cases by comparing the predicted heatmap against a binarized version of the ground-truth heatmap. **AP:** summarizes the Precision–Recall trade-off for the *inout* binary classification. $\mathbf{P_{LAH}}$: evaluates the precision of looking at people's heads. **OGA:** measures the accuracy of the gaze point falling inside the target object's bounding box. **SGA:** computes the accuracy of the target object's semantic class.

4.1.1 Distance

Distance is arguably the most widely used metric for evaluating gaze following models. It was first introduced in Recasens et al. (2015) and measures the Euclidean (L2) distance between the predicted and ground-truth gaze points on a 1×1 normalized square, yielding values in the range $[0, \sqrt{2}]$. Because the GazeFollow test set provides multiple gaze annotations per person, the protocol adapts the metric in two ways:

- **Minimum Distance:** the distance between the prediction and the closest ground-truth annotation, reflecting whether the model captures *any* plausible gaze target.
- **Average Distance:** the distance between the prediction and the mean of all annotations, reflecting agreement with the annotator *consensus*. While useful, this measure may become misleading in multi-modal cases, where averaging widely separated annotations produces a location that is not itself a valid target (Bao et al., 2022). This issue was further analyzed in our work (Tafasca et al., 2024b), though in practice we observed that both distance variants usually lead to similar conclusions under the standard heatmap-based formulation.

Although distance is typically computed in normalized image coordinates, some works also report raw pixel distances (Fan et al., 2018). However, this raises comparability issues when images or target objects differ in scale and resolution. A related metric from keypoint detection tasks is the Percentage of Correct Keypoints (PCK) (Munea et al., 2020), which counts predictions within a normalized distance threshold from the ground-truth. Unfortunately,

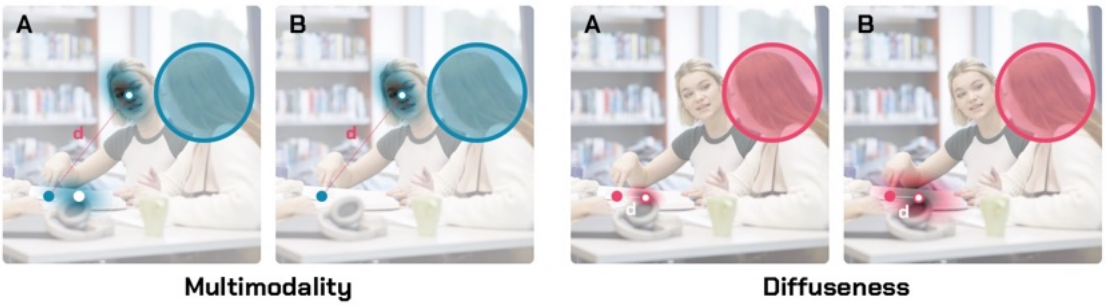


Figure 4.2: Limitations of the distance metric. **Multimodality:** Two example heatmap predictions with different numbers of modes. The ground-truth gaze point is shown in blue. Predicted modes are marked with white circles, and the argmax is highlighted with a blue border. Although both predictions have their argmax on the wrong target, prediction **A** is preferable because its secondary peak lies close to the ground truth. The distance metric, however, collapses the heatmap to a single point and assigns the same score to both. **Diffuseness:** Two monomodal predictions with different variances. The ground-truth gaze point is shown in pink, and the argmax in white with a pink border. Here, prediction **B** is preferable because its broader distribution covers the correct target, despite the argmax being slightly misplaced.

PCK cannot be directly applied to gaze following since it requires a meaningful per-sample normalizer (e.g. a target object bounding box or scale-dependent radius for "stuff" regions like walls or the sea), which is not consistently available.

4.1.2 Area Under the Curve (AUC)

While distance metrics evaluate a single predicted gaze point, they overlook two common properties of gaze heatmaps: *multimodality* (multiple plausible targets) and *diffuseness* (location uncertainty), as illustrated in Figure 4.2. To account for this, Recasens et al. (2015) introduced the AUC metric as part of the GazeFollow protocol. Here, the predicted heatmap is compared to a binary ground-truth heatmap that marks annotated gaze locations with 1 and all other pixels with 0. The task is framed as per-pixel binary classification, and the area under the ROC curve is computed. This measure is meaningful only when multiple annotations are available, which is the case for GazeFollow but not for VideoAttentionTarget or ChildPlay.

AUC, however, comes with notable caveats. Heatmaps are extremely imbalanced: the positive region around the gaze point is tiny compared to the large negative background. As a result, even broad or imprecise predictions often achieve high scores by ranking positives above most negatives. Once a heatmap is reasonably peaked near the correct point, nearly all positives are already ranked above negatives, so further gains in localization accuracy barely affect the score, leading to saturation. Thus, AUC is best interpreted as a coarse check of whether the model attends to the right *region*, but it is insufficient for assessing fine-grained accuracy. Similar limitations have been discussed in the saliency prediction literature (Bylinskii et al., 2018; Jia and Bruce, 2020). In practice, distance remains more interpretable and directly informative, as it quantifies pixel-level deviation from the annotation.

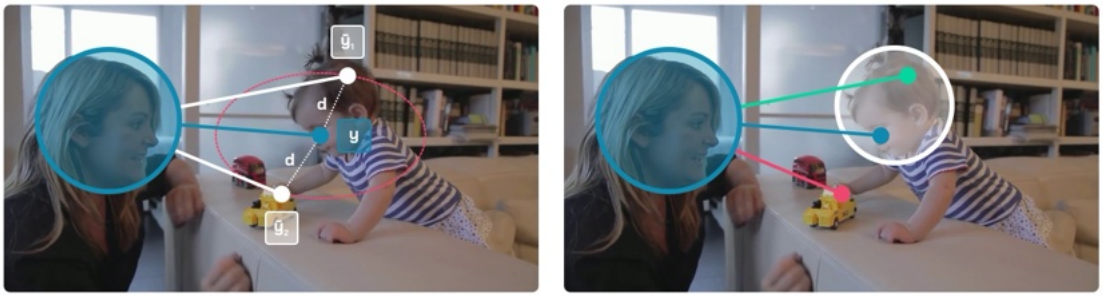


Figure 4.3: Left: Two predictions, $\bar{y}_1$ and $\bar{y}_2$, lie at the same distance d from the ground-truth point y. Should they be considered equally good? **Right:** Clearly not. The prediction $\bar{y}_1$ (in green) falls on the same object as the ground truth—the child's head—whereas $\bar{y}_2$ (in pink) falls on the toy. The Looking At Heads (LAH) metric addresses this by measuring the precision of predicting gaze toward people's heads, thereby adding a crucial semantic layer to evaluation that distance-based metrics alone cannot capture.

4.1.3 Average Precision (AP)

This metric evaluates the binary task of determining whether a subject's gaze falls inside the image boundaries or outside. It was first introduced in the VideoAttentionTarget protocol (Chong et al., 2020b) and later adopted in ChildPlay (Tafasca et al., 2023b). Average Precision is computed in the standard way, by integrating precision–recall curves over all thresholds.

4.1.4 Precision of LAH (P_{LAH})

As noted in Chapter 3, distance-based metrics conflate geometric proximity with semantic correctness: a prediction may lie close to the ground-truth point while still falling on the wrong object. To address this limitation, we introduced the first object-centric evaluation measure for gaze following (Tafasca et al., 2023b), marking a shift from dense to coarse and from geometric to semantic assessment. The P_{LAH} metric is designed to capture the social signal of looking at faces. A prediction is counted as a true positive if both the predicted and ground-truth gaze points fall inside *any* head bounding box. Conversely, a false positive is counted when the ground-truth gaze lies inside a head box but the prediction does not. We motivate our metric with an example in Figure 4.3.

4.1.5 Object Gaze Accuracy (OGA)

We introduced this metric in Tafasca et al. (2024a) to evaluate *localization* performance on GazeHOI, where annotations consist of bounding boxes around target objects rather than gaze points. A naive approach is to use the box center as a proxy gaze point, which may be reasonable for small objects but becomes unreliable for large ones. To address this, OGA measures whether the predicted gaze point falls inside the annotated object bounding box. In other words, it extends the object-centric idea of P_{LAH} from people's heads to arbitrary objects.

This metric captures the intuition that gaze following is ultimately about identifying the object

of attention, not just geometric proximity. Structurally, OGA resembles the Percentage of Correct Keypoints (PCK) but avoids the need for a scale-dependent normalizer. However, it is not a substitute for distance-based measures, since for large objects the precise location within the box (e.g. left vs. right side of a table or boat) still carries meaningful information.

4.1.6 Semantic Gaze Accuracy (SGA)

We introduced this metric in Tafasca et al. (2024a) to evaluate gaze target *recognition* performance on both GazeHOI and GazeFollow. SGA is defined as the top-k accuracy between predicted and ground-truth object class labels. In cluttered scenes, multiple plausible objects often lie near the gaze point, making top-3 accuracy more informative than strict top-1. For example, if a person is cutting a cake with a knife, the gaze may plausibly fall on either the cake or the knife, and both should be considered valid targets.

However, top-k accuracy has an important limitation: it does not guarantee that all entries among the top-k predictions are reasonable, only that at least one overlaps with the ground truth. To better handle this ambiguity, GazeFollow adopts a multi-label annotation protocol in which several objects can simultaneously be marked as gaze targets. In this case, we extend the standard definition to a set accuracy, where a prediction is counted correct if it belongs to the set of annotated ground-truth labels (e.g. either "knife" or "cake" would be correct if the ground-truth set is $\{\text{knife}, \text{cake}\}$).

4.2 Implementation Details

4.2.1 Resolution

Most two-stage gaze following methods operate on input images resized to 224×224, with some extending to 256×256. When used, head crops are also normalized to 224×224 to preserve fine-grained ocular cues such as eye direction. More recent approaches (Song et al., 2024; Ryan et al., 2025) have begun adopting higher resolutions (448×448 or 512×512), as the task is inherently a dense prediction problem and benefits from finer detail for accurately localizing small-scale targets.

In contrast, one-stage methods follow the original DETR resolution protocol, typically using image sizes between 800 and 1300 pixels depending on the aspect ratio. This higher resolution is crucial to compensate for the absence of explicit head crops and to ensure that faces remain large enough in the input for reliable gaze prediction.

4.2.2 Coordinate Encoding

Ground-truth gaze annotations typically consist of a single 2D point in the image. To train models that predict gaze as a spatial distribution, this point is converted into a heatmap by

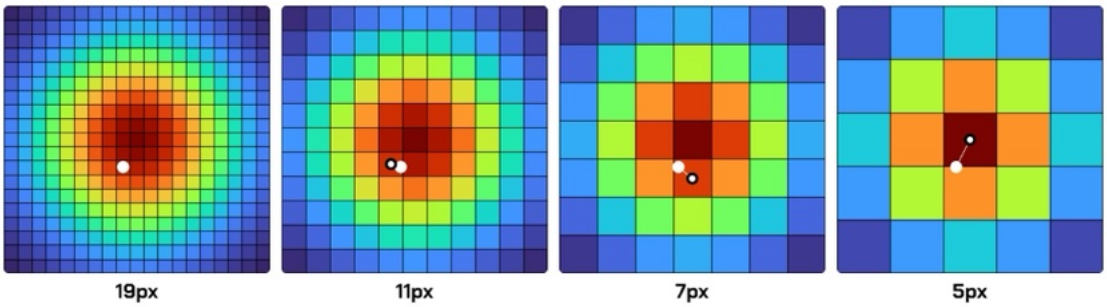


Figure 4.4: Illustration of the quantization problem in heatmap-based gaze prediction. A target point (white dot) lies exactly at the center of a pixel in a high-resolution heatmap (leftmost). When the same point is projected onto lower-resolution heatmaps, it must be snapped to the nearest discrete pixel center (white dot with black border). This discretization error grows as the resolution decreases, showing how coordinate decoding from coarse heatmaps can misalign the predicted gaze point even if the underlying model is accurate.

placing an isotropic Gaussian kernel centered at the annotation — a step often referred to as *coordinate encoding* in the literature. The standard practice is to downsample the image to a fixed resolution and generate the heatmap at a lower spatial resolution, both to reduce computational cost and to stabilize training. The kernel size (σ) is usually chosen relative to the heatmap resolution, ensuring that the annotated point occupies a small but visible region. In practice, all gaze following works use a heatmap resolution of 64×64 and a kernel size of 3.

While effective, this procedure introduces a quantization effect: fine differences between predictions are lost when projected onto a coarse heatmap grid (*cf.* Figure 4.4). This is especially problematic for small objects or targets close to each other, and has been discussed extensively in the literature (Lan et al., 2021b,a; Hu et al., 2024a). For this reason, decoding and evaluation often require additional care to avoid biasing performance.

4.2.3 Coordinate Decoding

Most gaze prediction models output a heatmap rather than a point directly. To evaluate predictions, the heatmap must be converted back into image coordinates. The standard approach is to take the argmax of the heatmap, which selects the pixel with the highest predicted probability. Some methods instead use a soft-argmax (Nibali et al., 2018; Luvizon et al., 2019; Jin et al., 2021), i.e., the expectation over the heatmap distribution. The main motivation for this choice is differentiability: unlike argmax, soft-argmax allows gradients to propagate through the decoding step, enabling end-to-end training directly on coordinates. A secondary advantage is that it provides sub-pixel precision, which can reduce quantization artifacts introduced by low-resolution heatmaps. This distinction matters particularly for dense evaluation metrics such as distance, since two models may produce visually different heatmaps but still yield the same argmax coordinate. Finally, decoding is performed relative to the heatmap resolution, so predictions must be rescaled back to the original image size. This step that can further amplify quantization issues when heatmaps are much smaller than the input image.

Beyond argmax and soft-argmax, several *post-processing* strategies have been proposed to obtain more accurate point estimates from heatmaps. In the pose estimation literature, a common heuristic is to refine the argmax by shifting the predicted coordinate by a small fraction (e.g. 0.25 pixels) in the direction of the second-highest activation (Newell et al., 2016), which partially compensates for discretization on the heatmap grid. More principled approaches have also been explored, such as DARK (Distribution-Aware coordinate Representation of Keypoints) by Zhang et al. (2020), which models the heatmap around the peak using a second-order Taylor expansion. This allows the method to adaptively account for the local curvature of the distribution and produce sub-pixel accurate predictions in a way that is both more robust and less heuristic-driven than simple offset corrections. While these techniques were originally proposed for pose estimation, the same principles apply to gaze prediction, where small improvements in decoding precision can meaningfully affect distance-based evaluation metrics. In our earlier works, we used the standard argmax but eventually transitioned to the DARK decoding scheme.

4.2.4 Data Augmentation

Data augmentation is critical in gaze following due to the limited diversity and imbalance of existing datasets. Most methods use standard augmentations such as random horizontal flipping (important because gaze direction is not inherently left–right biased), random cropping (to simulate different framing of the scene), and color jittering or blur (to improve robustness to lighting and video compression). In our work, we also apply head bounding box jittering, which perturbs the head crop during training and improves robustness to inaccuracies in the first-stage head detector. Augmentations must also be applied consistently to both the scene image and the corresponding head crop and gaze point, ensuring that geometric relations between them remain valid.

More advanced augmentations such as MixUp (Zhang et al., 2017) and CutMix (Yun et al., 2019a) are generally avoided in gaze following. These techniques combine parts of different images, which often destroys the spatial and social context that is essential for the task. For example, gaze depends on the relation between the head orientation and the surrounding scene; blending two unrelated scenes can break this link, creating ambiguous or misleading training signals. Similarly, mixing people across images can produce physically impossible interactions, which confuse the model rather than improve robustness. That being said, empirical validation of this hypothesis remains an open question.

4.2.5 Context People

Many benchmarks annotate only one person per image with a ground-truth gaze label, even though multiple people may be present. To better capture social cues and enable multi-person training, we introduced additional automatically detected heads as context people (Tafasca et al., 2024b; Gupta et al., 2024b). These are extracted using off-the-shelf head detectors,

filtered by confidence thresholds and by overlap with annotated subjects, and included in the model input alongside the annotated person. During training, however, supervision is applied only to the annotated person, while context people act as auxiliary cues. At inference time, the model can process all visible people jointly. This strategy improves efficiency by enabling multi-person architectures. We will explore this topic further in Chapter 6.

4.2.6 Loss Functions

The training objective combines three complementary components: a heatmap regression loss, a gaze vector loss, and an in/out classification loss. Each captures a different aspect of the gaze prediction problem.

Heatmap Regression Loss. The primary supervision signal is the predicted gaze heatmap $\hat{H} \in \mathbb{R}^{W \times H}$, trained to match a Gaussian-centered ground-truth heatmap H. The most common choice is a pixel-wise Mean Squared Error (MSE):

$$\mathscr{L}_{\text{hm}} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \left(H(x,y) - \hat{H}(x,y) \right)^2. \tag{4.1}$$

Alternatively, some works (Ryan et al., 2025) employ Binary Cross-Entropy (BCE):

$$\mathscr{L}_{\text{hm}}^{\text{BCE}} = -\frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \Big[H(x,y) \log \hat{H}(x,y) + (1 - H(x,y)) \log(1 - \hat{H}(x,y)) \Big]. \tag{4.2}$$

In keypoint prediction with Gaussian heatmaps, MSE treats the target as a continuous probability-like distribution and encourages the model to reproduce its shape, yielding smooth, calibrated peaks where intensity decays with distance from the keypoint. BCE, in contrast, interprets each pixel as an independent Bernoulli classification problem (i.e. near-peak pixels are "positive," others are "negative"), so the model is only incentivized to separate plausible from implausible pixels, often producing broad, *flat plateaus* around the target. On the other hand, BCE tends to train faster because it provides stronger gradients when predictions are wrong, especially on background pixels, whereas MSE gradients are weak away from the peak. In practice, MSE is more suitable when calibrated, distributional heatmaps or uncertainty estimates are important, whereas BCE can suffice if only the argmax coordinate matters. This distinction reflects a *generative* vs. *discriminative* view of the problem. Empirically, both yield similar performance in gaze following, but MSE is generally preferred because it produces better distributions. To illustrate these differences, we compare the predicted heatmaps of two networks trained on a toy keypoint detection problem using MSE and BCE (*cf.* Figure 4.5).

Gaze Vector Loss. To encourage geometric consistency with head orientation, the model also predicts a normalized 2D or 3D gaze direction vector $\hat{g} \in \mathbb{R}^d, d \in \{2,3\}$, supervised against the

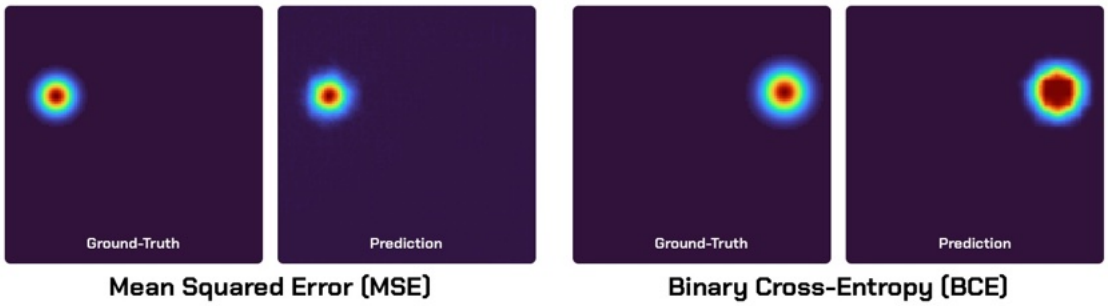


Figure 4.5: Comparison of predicted heatmaps when training with different losses. **MSE:** The prediction decays smoothly from the peak and matches the intermediate values of the ground truth, yielding a soft, calibrated, probability-like shape. However, faint residuals remain in the background because the MSE gradient in low-value regions is very small, providing little incentive to push predictions exactly to zero. **BCE:** The prediction preserves a round overall shape but lacks calibration: a broad central area is assigned nearly uniform peak intensity before dropping sharply to zero. BCE produces cleaner, mask-like heatmaps without residuals since background errors receive stronger gradients, driving values closer to zero.

ground-truth vector g. We adopt a cosine similarity loss:

$$\mathscr{L}_{\text{vec}} = 1 - \frac{\langle g, \hat{g} \rangle}{\|g\|_2 \, \|\hat{g}\|_2}. \tag{4.3}$$

This formulation minimizes the angular deviation between predicted and ground-truth gaze directions. Importantly, we avoid using the explicit arccosine of the angle, as its gradient is not defined everywhere and can lead to numerical instabilities during training.

In/Out Classification Loss. In addition, the model predicts whether the gaze target lies inside or outside the image boundaries. This binary decision is supervised with a standard Binary Cross-Entropy loss:

$$\mathscr{L}_{\text{inout}} = -\Big[y \log \hat{p} + (1-y) \log(1-\hat{p}) \Big], \tag{4.4}$$

where $y \in \{0,1\}$ is the ground-truth label and $\hat{p}$ is the predicted probability of the gaze being inside the image.

Total Loss. The final objective is a weighted combination of the three terms:

$$\mathscr{L} = \lambda_{\text{hm}} \mathscr{L}_{\text{hm}} + \lambda_{\text{vec}} \mathscr{L}_{\text{vec}} + \lambda_{\text{inout}} \mathscr{L}_{\text{inout}}, \tag{4.5}$$

with weights $\lambda_{\text{hm}}, \lambda_{\text{vec}}, \lambda_{\text{in/out}}$ set to balance the contribution of each loss.

4.2.7 Training Hyperparameters

Most of our models are trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with weight decay, combined with a cosine annealing learning rate schedule and warmup. Sometimes we also use Stochastic Weight Averaging (SWA) (Izmailov et al., 2018) to improve

generalization. Training on GazeFollow is performed for 20 epochs on a single NVIDIA RTX 3090 GPU, which typically requires 10–16 hours depending on model size and the number of context people. For video-based datasets such as VideoAttentionTarget and ChildPlay, we only perform lightweight fine-tuning for 2-10 epochs, which is substantially faster, usually taking 1–6 hours per experiment.

4.3 Training and Evaluation Routines

GazeFollow. We train on GazeFollow from scratch, often initializing the backbone with weights pretrained on large-scale image datasets such as ImageNet (Deng et al., 2009) or Gaze360 (Kellnhofer et al., 2019), while connector blocks and task-specific layers are initialized randomly. Training is performed at the image level on a single person at a time (i.e. an image with two annotated people will contribute two independent samples). GazeFollow serves as the main source of pretraining due to its large scale and diversity, and provides the initialization point for fine-tuning on other datasets.

VideoAttentionTarget. For VAT, training is performed in a static and per-person manner: a sample corresponds to a single annotated person, not an entire frame, since multiple people may be annotated in the same frame. Frames are downsampled with a temporal stride of 5, either by keeping one every 5 frames or by sampling randomly from each group of 5, which reduces redundancy while maintaining temporal coverage. We always fine-tune on VAT starting from a model pretrained on GazeFollow, optionally freezing shared components where appropriate and training only the task-specific heads for a few epochs to avoid overfitting. In addition, we sometimes report cross-dataset evaluation by applying the pretrained GazeFollow model directly to the VAT test set without fine-tuning, which highlights generalization across static and video-based domains.

ChildPlay. The training protocol for ChildPlay mirrors that of VAT, including static per-person sampling and a temporal stride of 5. However, we restrict supervision to the binary in/out label and the gaze point annotations, discarding additional classes such as gaze shifts, eyes-closed, or occluded targets. As with VAT, fine-tuning is always performed from a GazeFollow-pretrained model, and cross-dataset evaluation may also be reported. This ensures that the model focuses on consistent gaze-following behavior while ignoring noisy or ambiguous cases.

GazeHOI. Although GazeHOI is image-based and relatively large, we also pretrain on GazeFollow before fine-tuning on its training set. This is because GazeHOI only provides bounding boxes for gaze objects and does not include explicit gaze points. By default, we use the center of the object box as a proxy gaze point; however, this can be problematic for large objects. For example, the center of a person's bounding box often falls near the hips, whereas the actual gaze target may lie on the head. To mitigate this issue, we adopt a hybrid strategy: if the object box exceeds a certain size threshold, we use the prediction from a pretrained GazeFollow model as the pseudo ground-truth, provided that this prediction lies within the annotated

object box. Otherwise, we fall back to the box center. This approach better preserves semantic consistency between gaze points and object-level targets.

Model Selection. For all datasets, the best checkpoint is selected based on the *distance metric* computed on the validation set. This ensures consistency and aligns model selection with the most widely used evaluation criterion in gaze following.

4.4 On the Limitation of Video-based Benchmarks

Despite their natural appeal for modeling dynamic behaviors and temporal events, video-based datasets for gaze following suffer from intrinsic limitations. Most notably, they provide large numbers of frames but very little independent variability. In order to investigate this topic further, we studied the distributional differences between GazeFollow (GF), VideoAttentionTarget (VAT), and ChildPlay (CP).

4.4.1 Qualitative Analysis

Setup. We extract DINOv2 embeddings (Oquab et al., 2023) for all datasets and sub-sample to maintain readability and reduce visual clutter: **GF** (50k images), **VAT** (2 frames/clip, ~1.2k), and **CP** (2 frames/clip, ~0.8k). We also sample 100k images from *two additional static datasets* used only as *anchors* to stabilize the manifold (they are not rendered in the plot). A strong down-sampling of the video sets avoids clip-level micro-clusters. We fit a single UMAP (McInnes et al., 2018) (Euclidean metric) on the union and project to 2D.

Results. The resulting layout (Figure 4.6) shows **GF** spanning a broad region of the visual manifold, while **VAT** and **CP** form compact islands that lie *inside* GF's span. The similarity of prototypical images derived from GF and the video datasets further substantiates this observation. We also note that including the anchor datasets does not change this qualitative picture; it simply makes the geometry less sample-idiosyncratic.

4.4.2 Quantitative Analysis

We measure containment and variability directly in the original feature space (not UMAP) using nearest-neighbor distances.

Setup. Let $d_k(x, S)$ denote the distance from a point x to its k-th nearest neighbor in the set S. For a reference dataset A, define the leave-one-out self-distances

$$\mathscr{D}_{A\to A} = \{d_k(a, A \setminus \{a\}) : a \in A\}.$$

Let $\tau_q(A)$ be the q-th quantile of $\mathscr{D}_{A\to A}$ (with $q \in [0,1]$). For a second dataset B, define the

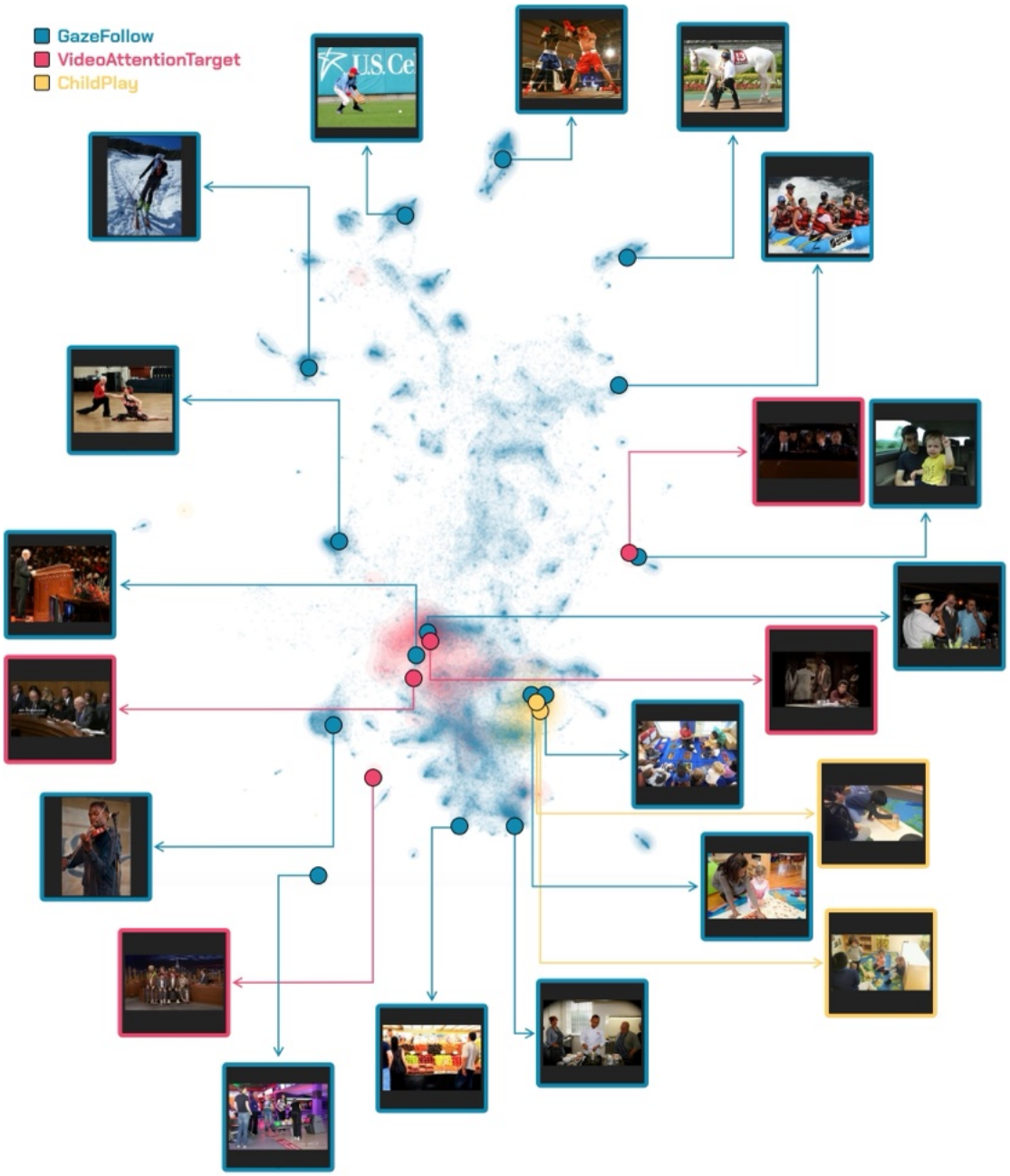


Figure 4.6: UMAP visualization of DinoV2 image embeddings from three gaze datasets: GazeFollow, VideoAttentionTarget, and ChildPlay. Each point corresponds to a single image, and representative samples are shown around the embedding. GazeFollow, a large and diverse static image dataset, covers a wide span of the manifold, while video-based datasets (VideoAttentionTarget, ChildPlay) occupy much smaller and denser regions that lie almost entirely within the coverage of GazeFollow. This illustrates the limited variability and strong redundancy of video datasets, where many near-duplicate frames collapse into compact clusters, underscoring why large-scale static datasets are critical for pretraining gaze models before fine-tuning on smaller video collections.

coverage curve of B relative to A as

$$\mathrm{Cov}_q(B \to A) \;=\; \frac{1}{|B|} \sum_{b \in B} \mathbf{1}\big[d_k(b, A) \le \tau_q(A)\big],$$

i.e., the fraction of B whose k-NN in A is no farther than a typical within-A threshold at quantile q. We use $k=1$ and report the scalar

$$\mathrm{Coverage@95}(B \to A) \;=\; \mathrm{Cov}_{0.95}(B \to A).$$

We also plot histograms of (*i*) $A \to A$ and (*ii*) $B \to A$ 1-NN distances, and, when assessing within-set variability, $B \to B$ (leave-one-out) distances.

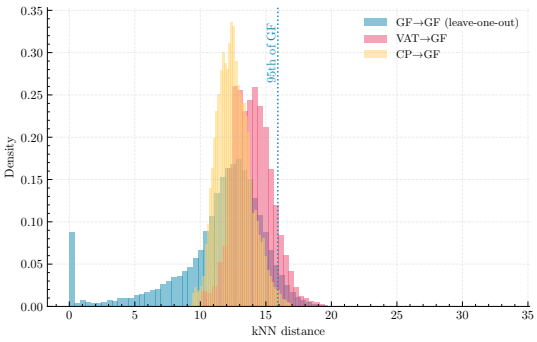
Results. For **VAT→GF** and **CP→GF**, the coverage curves rise steeply and saturate near 1, with **Coverage@95** of approximately **0.90** and **0.99**, respectively (Figs. 4.7b and 4.7d). This indicates that almost all frames in the video datasets are as close to some GF image as the 95th-percentile *within-GF* neighbor distance; i.e. *VAT* and *CP* are largely contained in GF. In contrast, **VAT→CP** exhibits essentially zero coverage, confirming that the two video datasets occupy largely disjoint subregions unless GF is used as a bridge (consistent with their sources: childcare centers for CP vs. TV shows for VAT).

Interpretation via distance histograms. For both **VAT→GF** and **CP→GF**, the distance distributions lie well within GF's self-distance range (Figure 4.7a), again supporting the containment hypothesis: *VAT, CP* $\subset$ *GF*. Crucially, the **CP→CP** histogram is *narrow*, whereas **GF→GF** is *broad* (Figures 4.7a and 4.7c). This directly evidences CP's low within-set variability compared to GF (and similarly for VAT). Together, these results show that video datasets contribute little *distributional novelty* beyond GF and contain many near-duplicate frames (low effective sample size), despite large raw frame counts.

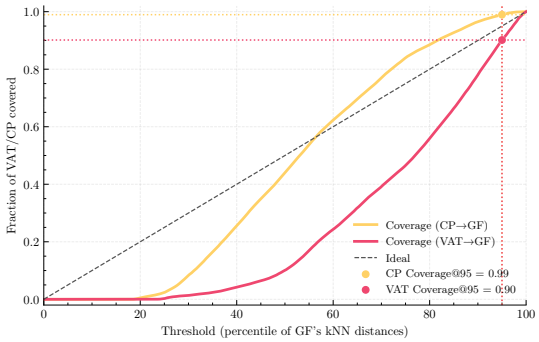
For gaze following, diversity is critical because solving the task depends on the interplay of three components:

- **Saliency detection:** highlighting candidate objects and regions that might attract attention
- **Geometric constraints:** estimating the person's field of view from head pose or eye direction
- **Target selection:** resolving which salient candidate in the gaze cone is most likely the actual target

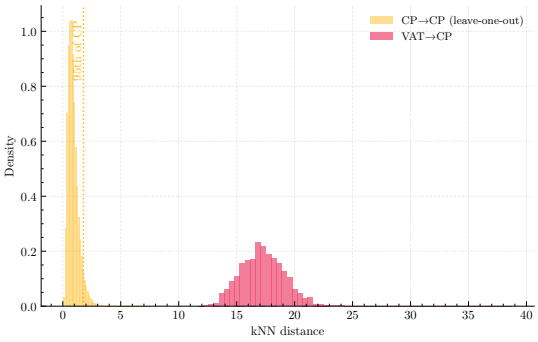
While video datasets do contain variability in terms of gaze directions (people look in many directions within a clip), what they lack is variability in scene content, object categories, and social contexts. Without seeing a wide range of environments, people, and objects, models cannot reliably learn which elements are plausible gaze targets. For example, learning that a ball in a stadium is often the focus of attention, or that another person's face is a common target, requires exposure to diverse contexts, object types and annotation statistics.



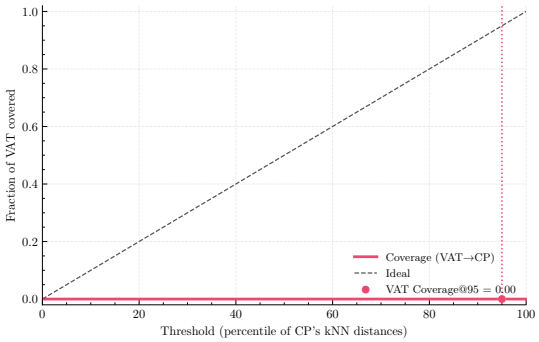
(a) Histogram of k-NN distances: VAT→GF and CP→GF



(b) Coverage curves: VAT→GF and CP→GF



(c) Histogram of k-NN distances: VAT→CP



(d) Coverage curve: VAT→CP

Figure 4.7: Comparison of dataset distributions in DINO feature space. Top row: VideoAttentionTarget (VAT) and ChildPlay (CP) relative to GazeFollow (GF), shown via k-NN distance histograms (left) and coverage curves (right). Bottom row: VAT relative to CP, shown via histogram (left) and coverage curve (right).

These associations cannot emerge from narrow video datasets with limited scene and object coverage.

This explains why the standard experimental protocol in gaze following is to pre-train on a large, diverse static dataset such as GazeFollow, which provides broad coverage of natural scenes, people, and object categories, before fine-tuning on video datasets (Chong et al., 2020b). In practice, our findings confirm that models trained only on video data overfit to repetitive frames and perform poorly, whereas models pre-trained on GazeFollow generalize better and often outperform video-trained models even when evaluated on the video benchmarks themselves.

Methods Part II

5 Multimodality & Scene Geometry

Images are flat; gaze is not. Predicting where someone looks from a single image is deceptively hard because the world that generates that 2D image lives in a 3D space. A person looking toward a cup on the table might align, by pure projection, with the lamp on the wall behind it. In 2D, both are collinear along the same ray, yet only one is physically reachable and visually plausible. Depth, occlusion, and field-of-view constraints are the missing pieces that shift the inference logic from appearance-based saliency to spatial reasoning.

This chapter opens Part II (Methods) of this manuscript by putting *geometry and multimodality* at the heart of gaze prediction. We focus on a depth-aware *3D Field-of-View (3D FoV)* representation that aligns a person-specific 3D gaze direction with the scene's inferred geometry to produce a physically grounded gaze heatmap and in/out decision (Tafasca et al., 2023b). The design stands in contrast to classic 2D formulations and to our earlier multimodal baseline (Gupta et al., 2022), which improved robustness but left geometric consistency implicit. Here, we make geometry explicit and show how doing so resolves projection ambiguities, improves generalization, and provides clearer failure modes. We illustrate the difference between 2D and 3D field-of-view heatmaps in Figure 5.1.

5.1 Motivation

At its core, gaze following in 2D suffers from projection ambiguity: many distinct 3D targets map to the same 2D image ray. Traditional two-branch CNNs learn to prefer some of these candidates using appearance and coarse directional cues, but when multiple plausible objects lie along the line of sight, or when background structure is richly textured, they cannot enforce physical plausibility. Three issues arise:

- **Line-of-sight conflation:** In 2D formulations, the gaze direction is represented as a line or cone in image space that illuminates all pixels collinear with the head's orientation. As a result, distant background regions along the same projection ray appear equally plausible, even though they lie at very different depths.

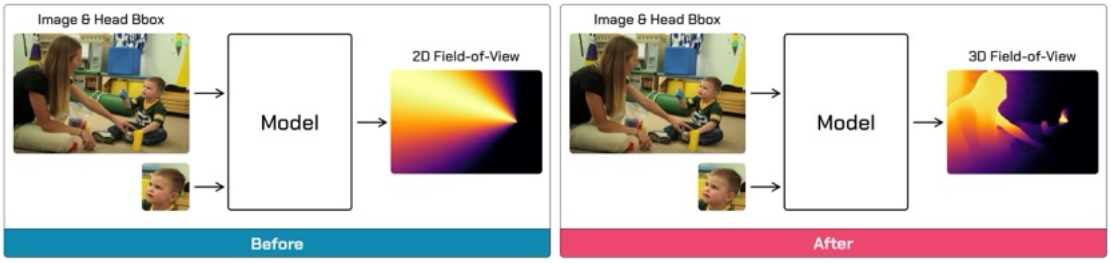


Figure 5.1: Illustration of the difference between 2D and 3D field-of-view representations. Unlike the 2D FoV, which highlights all points along the image ray, the 3D FoV respects scene depth and excludes background regions that are not geometrically consistent with the gaze direction.

- **Scale and perspective:** The same pixel offset corresponds to different 3D gaze angles depending on depth, yet 2D supervision penalizes them identically. The model therefore learns direction in image coordinates rather than in physical space and cannot distinguish small angular deviations on distant surfaces from large ones on nearby objects.
- **Geometric plausibility:** Without depth, 2D models cannot distinguish between points that merely project along the gaze ray and those that are actually visible in 3D space. Depth provides this missing spatial constraint, ruling out targets hidden behind obstacles and enforcing coherence with the scene's physical structure.

Depth reintroduces the missing constraints. A monocular depth map, converted into a pseudo point cloud in the camera frame, lets us move from "which pixels align with the 2D gaze ray?" to "which 3D points lie within a plausible field of view from the head's position and gaze direction?" Operationally, our model *GeomGaze* (i) predicts a normalized 3D gaze vector from the head crop; (ii) constructs a 3D FoV representation by measuring directional agreement (e.g. cosine similarity) between this vector and 3D scene points, modulated by the gaze cone's aperture; and (iii) projects this physically filtered distribution back into the image to guide the gaze heatmap and in/out score. This pipeline turns depth from an auxiliary feature into a spatial gate, pruning impossible regions and sharpening plausible ones.

This geometric grounding also re-imagines supervision. While datasets annotate 2D gaze points, they rarely provide 3D gaze vectors. We therefore derive *pseudo-ground-truth 3D directions* by mapping the annotated 2D gaze point and head box center point to the camera frame, and computing a normalized vector between them.

Finally, this chapter positions our earlier multimodal 2D FoV baseline (Gupta et al., 2022) as a stepping stone. That model concatenated RGB, pose, and depth features into a conventional decoder and showed that multimodality helps. Yet it left the physics of visibility and 3D reachability to the network to infer implicitly. The 3D FoV representation we develop here makes the spatial constraints explicit, yielding cleaner inductive bias, more interpretable maps, and stronger cross-dataset transfer.

In summary, this chapter argues that geometry is not a luxury add-on but a prerequisite for reliable gaze prediction. By elevating depth from feature to representation via a 3D FoV that

couples a predicted 3D gaze direction with the scene's 3D structure, we resolve line-of-sight ambiguities and lay a principled foundation for 3D gaze following.

5.2 Related Work

5.2.1 Multimodal Fusion

One way to enrich gaze prediction is to leverage additional modalities that provide complementary context to the visual scene. Beyond depth, which we discuss in the next section, researchers have explored the use of body pose (e.g. Wei et al. (2018); Guan et al. (2019); Bao et al. (2022); Toaiari et al. (2024); Yang and Lu (2024)), audio or speech cues (e.g. Hou et al. (2024)), and object or activity representations (e.g. Wei et al. (2018); Guan et al. (2019); Hu et al. (2022b); Tonini et al. (2023); Yang and Lu (2024)). These signals can supply information that is either missing or ambiguous in the RGB domain: pose constrains the direction of attention when the eyes are not visible; depth informs geometric reachability; and semantic or auditory context narrows the range of plausible gaze targets.

In our earlier work Gupta et al. (2022), we proposed a modular multimodal architecture designed to integrate such heterogeneous cues in a scalable and interpretable way. Each modality (i.e. RGB appearance, monocular depth, and body pose rendered on a blank canvas) was processed by an independent convolutional branch with a shared latent dimensionality, allowing the model to learn both modality-specific representations and cross-modal interactions through late fusion. This modular design offered two key advantages: (i) it allowed flexible inclusion or removal of modalities at training and inference time, and (ii) it encouraged complementary reasoning, where pose provided directional priors, depth grounded the geometry of the scene, and RGB captured semantic and contextual cues. In particular, combining pose and depth was important: pose not only anchors the orientation of the gaze but also provides explicit spatial localization of people in the scene (who themselves are frequent gaze targets), while depth constrains the spatial extent of that orientation (*cf.* Figure 5.2). Together, these cues improving disambiguation in cluttered or low-resolution settings where appearance alone is unreliable.

Unlike works that directly concatenate multimodal features into a monolithic encoder, our formulation maintained separable processing pipelines and combined them only at intermediate or decision levels. This design improves interpretability and robustness to missing modalities, a property that becomes particularly valuable in real-world or privacy-sensitive applications. We revisit this topic in Chapter 10.

5.2.2 Geometric Reasoning

Incorporating depth to reason about scene geometry is not new in gaze target prediction. Early approaches often treated depth as just another visual channel rather than a true geometric cue.



Figure 5.2: Example scenes where depth and pose provide useful cues for gaze target inference. **Left:** body pose reveals interaction between people, while depth helps disregard salient background objects. **Middle:** manipulation activities in which knowledge of hand position aids reasoning about gaze and attention. **Right:** depth information allows the model to filter out distant face candidates in the background..

For example, many works (e.g. Tonini et al. (2022); Gupta et al. (2022); Miao et al. (2023)) feed the depth map as an additional feature map concatenated with RGB or head inputs. While this can improve robustness, it does not impose any explicit geometric reasoning: depth is processed as an image-like signal rather than as a spatial structure.

A more sophisticated alternative is to use depth as supervision rather than input. Jin et al. (2022) follow this direction by distilling geometric information from a depth estimation branch into the gaze network. Intermediate depth features are fused into the scene pathway before decoding, allowing the model to benefit from geometric cues during training while remaining RGB-only at inference. Conceptually, however, this still treats depth as appearance projected in image space, rather than reasoning directly in 3D.

A few methods attempt to move beyond this limitation. Bao et al. (2022), for instance, reconstruct a coarse point cloud from the scene and refine it by using detected humans as reference objects. However, they never predict an explicit 3D gaze vector, and thus stop short of aligning the predicted gaze direction with the scene geometry. Other works introduced before and after our approach propose heuristic mechanisms to relate gaze and depth but lack a principled formulation. Fang et al. (2021), for example, use a monocular depth estimator to segment the scene into three zones (foreground, mid-ground, and background) based on relative depth to the subject. Similarly, Liu et al. (2024) and Yang and Lu (2024) multiply a 2D gaze cone with a proximity map derived from relative depth, effectively weighting candidate targets by distance but still operating in image coordinates.

Even among methods that nominally operate in 3D (e.g. Tonini et al. (2023)), many define the 3D scene in a pseudo-3D coordinate system: they take the image-plane pixel coordinates as (x, y) and append the predicted depth as z. This creates an inconsistent spatial representation, since the (x, y, z) values no longer correspond to metric coordinates in the camera frame: the depth varies with focal length and field of view, but the image-plane distances do not. As a result, geometric relations such as angular alignment or Euclidean distance become distorted, limiting the model's ability to reason about physical visibility and spatial coherence.

In contrast, our approach performs reasoning directly in true 3D space. We reconstruct a point cloud in the camera coordinate frame and define a 3D field of view (3D FoV) by measuring the directional agreement between the predicted 3D gaze vector and scene points within that space. The resulting geometric field forms a physically consistent 2D heatmap. Crucially, the entire model, including the gaze pathway and the 3D FoV construction, is trained end-to-end, ensuring that geometric reasoning is learned jointly rather than applied as a post-hoc operation.

5.3 Architecture

Figure 5.3 provides an overview of the proposed architecture. It follows a modular design that mirrors the human visual process: a *Gaze Encoder* extracts appearance-based cues from the head crop, a *Scene Encoder* analyzes the broader visual and geometric context, and a *3D Field-of-View representation* bridges the two by embedding geometric consistency between the predicted gaze direction and the reconstructed scene. Together, these components enable joint reasoning about where a person is looking and whether the gaze target lies inside or outside the image frame.

5.3.1 Gaze Encoder

The Gaze Encoder aims to predict the 3D gaze direction defined in the *local eye coordinate system* $\mathbf{C}_e$ associated with the head crop $\mathbf{h}_{\text{crop}} \in \mathbb{R}^{h \times w \times C}$. We express the gaze vector in $\mathbf{C}_e$ rather than in the camera coordinate system $\mathbf{C}_c$, as gaze direction depends primarily on *appearance-based cues* (e.g. head orientation, eye configuration, and facial geometry), rather than on the head's absolute position within the image.

In a camera-centric frame, the same face or eye appearance could correspond to very different gaze vectors depending on where the head appears in the image, making the mapping from pixel appearance to 3D direction non-stationary across the frame. By contrast, defining gaze in an *eye-centric* frame normalizes out global translation and scale, enabling the network to learn a consistent, person-independent relationship between local facial appearance and gaze direction. This formulation mirrors human perception: we infer where someone looks from the relative rotation of their eyes and head, not from their 2D screen position.

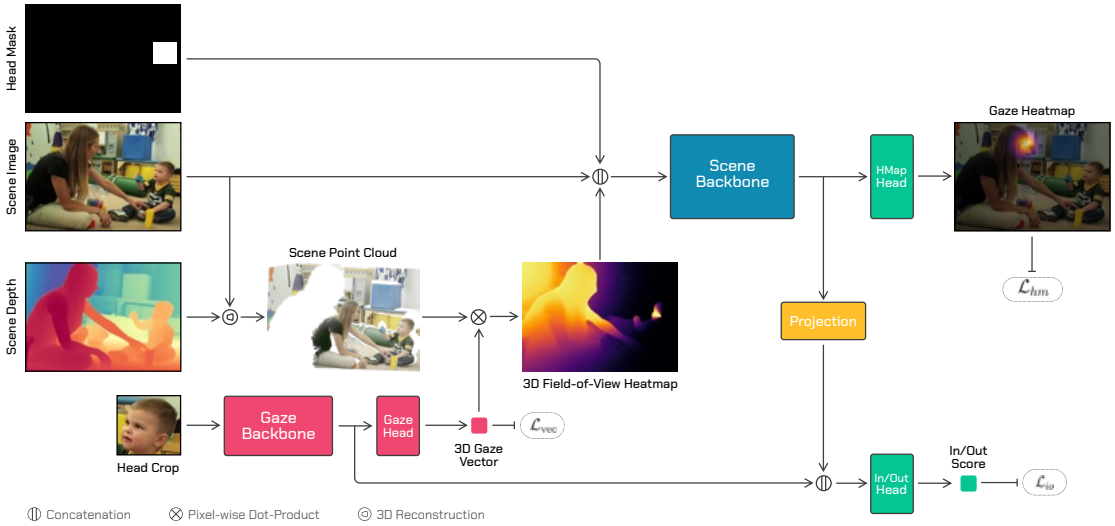


Figure 5.3: Overview of the proposed architecture *GeomGaze*. The model consists of two complementary pathways: a Gaze Pathway and a Scene Pathway. The gaze pathway encodes the head crop to produce a 3D gaze vector (in the eye coordinate system) and a latent gaze embedding. The scene pathway combines the RGB image, head location mask, and the depth-based 3D Field-of-View (FoV) heatmap derived from the gaze vector and point cloud. A shared scene backbone produces a person-conditioned scene saliency feature map, which is decoded into a 2D visual attention heatmap representing the predicted gaze location. The same feature map is also compressed into a 1D descriptor, concatenated with the gaze embedding, and fed to an MLP to estimate the in/out probability. All components are trained jointly with three supervision signals: gaze direction, heatmap localization, and in/out classification losses.

Formally, a gaze backbone $\mathcal{G}_b$ encodes the head crop into a latent embedding $\mathbf{x}_{\text{emb}} \in \mathbb{R}^d$, and a prediction head $\mathcal{G}_h$ with a `tanh` activation projects it onto the unit sphere to obtain a normalized gaze vector $\mathbf{g}_v \in \mathbb{R}^3$:

$$\mathbf{x}_{\text{emb}} = \mathcal{G}_b(\mathbf{h}_{\text{crop}}), \qquad \mathbf{g}_v = \mathcal{G}_h(\mathbf{x}_{\text{emb}}). \tag{5.1}$$

This vector represents the gaze direction in $\mathbf{C}_e$, the same coordinate system in which the point cloud will later be expressed. Finally, the intermediate embedding $\mathbf{x}_{\text{emb}}$ is also reused in the decoding stage to infer the in/out flag.

Eye Coordinate System. We define the eye coordinate system $\mathbf{C}_e$ as a local 3D frame centered at the midpoint of the subject's head bounding box, denoted $E_c \in \mathbb{R}^3$, which serves as an approximate head- or eye-center, expressed in the camera frame. This choice provides a stable and reproducible reference point across subjects and poses, even if it does not coincide with the actual eye location.

The z-axis of $\mathbf{C}_e$ is defined as the unit vector from the camera center $C_c = (0,0,0)$ to the head center:

$$\hat{\mathbf{z}}_e = \frac{E_c - C_c}{\|E_c - C_c\|} = \frac{E_c}{\|E_c\|}. \tag{5.2}$$

This axis represents the forward viewing direction from the camera toward the person. To resolve the rotation ambiguity around $\hat{\mathbf{z}}_e$, we use the camera's global up vector $\mathbf{u}_c = (0,1,0)^\top$ as a reference orientation. The x- and y-axes are then obtained by enforcing orthogonality using the cross product:

$$\hat{\mathbf{x}}_e = \frac{\mathbf{u}_c \times \hat{\mathbf{z}}_e}{\|\mathbf{u}_c \times \hat{\mathbf{z}}_e\|}, \qquad \hat{\mathbf{y}}_e = \hat{\mathbf{z}}_e \times \hat{\mathbf{x}}_e. \tag{5.3}$$

This construction yields a right-handed, head-centered frame whose orientation is consistent across images. The rotation matrix from the camera to the eye coordinate system is thus

$$\mathbf{R}_{c\to e} = \begin{bmatrix} \hat{\mathbf{x}}_e^\top \\ \hat{\mathbf{y}}_e^\top \\ \hat{\mathbf{z}}_e^\top \end{bmatrix}, \tag{5.4}$$

and the transformation of a 3D point P_c from the camera frame $\mathbf{C}_c$ to the eye frame $\mathbf{C}_e$ is given by

$$P_e = \mathbf{R}_{c\to e}\,(P_c - E_c). \tag{5.5}$$

This coordinate system follows the convention of Gaze360 (Kellnhofer et al., 2019) from the gaze estimation literature and ensures that both the predicted gaze vector and the reconstructed scene are expressed in a geometrically consistent local reference frame. Refer to Figure 5.4 for an illustration of this coordinate system.

5.3.2 3D Field-of-View Representation

The goal of the 3D field-of-view (3D FoV) representation is to geometrically align the predicted gaze vector with the reconstructed scene, producing a spatial activation map that highlights regions lying within the subject's plausible line of sight. This process involves two main steps: (i) reconstruction of a 2.5D scene from the RGB image and its corresponding depth map, and (ii) generation of a 3D field-of-view heatmap by matching the predicted gaze vector with the reconstructed scene.

3D Scene Reconstruction. Given the RGB scene image $\mathbf{I}$ and its monocular depth map $\mathbf{D}$, we reconstruct a point map in the camera coordinate system $\mathbf{C}_c$. Following the standard pinhole camera model, each pixel (u,v) with depth value $D(u,v) = Z_c$ is back-projected to the 3D space as:

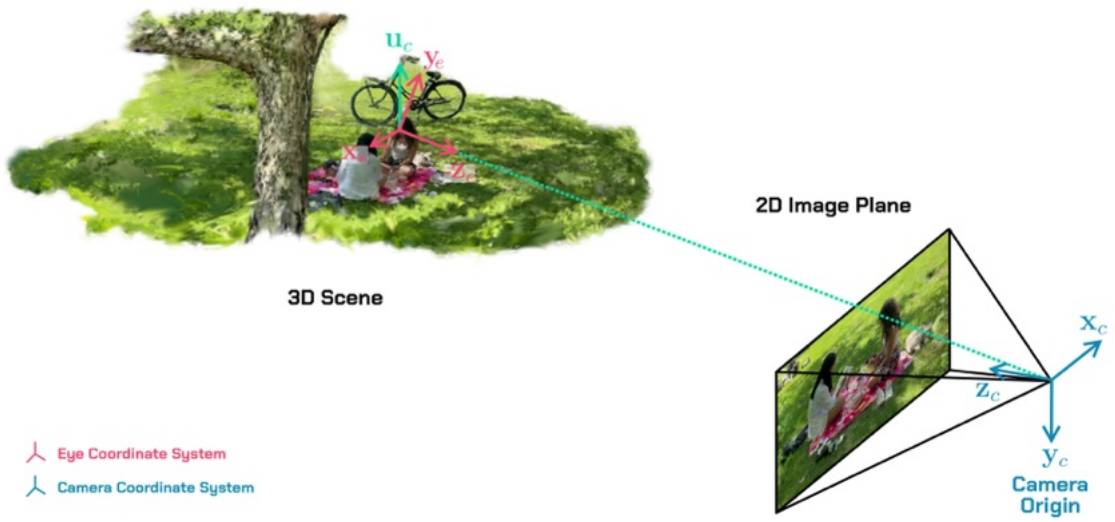


Figure 5.4: Illustration of the local eye coordinate system $\mathbf{C}_e$ for a given person, showing its relationship to the camera coordinate system $\mathbf{C}_c$ and the global up vector $\mathbf{u}_c$.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = D(u,v) \begin{bmatrix} \frac{u-c_x}{f_x} \\ \frac{v-c_y}{f_y} \\ 1 \end{bmatrix}, \tag{5.6}$$

where (f_x, f_y) are the focal lengths in pixel units and (c_x, c_y) denote the coordinates of the principal point. In practice, we assume *square pixels* ($f_x = f_y = f$), a *centered principal point* ($c_x = W/2$, $c_y = H/2$), and *no skew*, which are standard approximations when camera intrinsics are unknown. Thus, the only missing parameter to build the camera matrix is the focal length f, which we estimate from the scene image using the model from Yin et al. (2021).

This reconstruction is technically called a *2.5D point map*, meaning that each pixel is back-projected into a single 3D point along its viewing ray. Unlike a full 3D point cloud, which provides a complete set of (X, Y, Z) coordinates that describe the scene from all viewpoints, a 2.5D representation captures only the visible surfaces from a single perspective. Occluded or hidden regions are therefore absent, but this partial reconstruction remains sufficient for aligning the predicted gaze direction with the observed scene geometry. To build this point cloud, we use the depth estimation model from Patakin et al. (2022), which is designed to recover the shift parameter and preserve geometric consistency, leading to improved fidelity of reconstructed scene.

To ensure alignment between the gaze vector and the scene, the point map is transformed from the camera coordinate system $\mathbf{C}_c$ to the local eye coordinate system $\mathbf{C}_e$ using Equation 5.5. All subsequent reasoning, including gaze–scene matching, is performed in this common coordinate frame.

3D Field-of-View Heatmap Generation. Let $\mathbf{g}_v \in \mathbb{R}^3$ denote the normalized gaze vector

predicted by the Gaze Encoder, and let $\mathbf{p}_i$ denote the 3D point corresponding to pixel i in the reconstructed scene, both expressed in the eye coordinate system $\mathbf{C}_e$. For each point, we measure its *directional agreement* with the predicted gaze vector using cosine similarity:

$$c_i = \frac{\mathbf{g}_v \cdot \mathbf{p}_i}{\|\mathbf{p}_i\|}. \tag{5.7}$$

This value captures how well each 3D point lies along the predicted line of sight. A higher c_i indicates stronger alignment (smaller angular deviation). We apply an exponential decay to emphasize regions close to the gaze axis and suppress off-axis areas more sharply:

$$V_i = \begin{cases} c_i, & \text{if } c_i > \tau, \\ \tau \times \frac{\exp(\kappa c_i)}{\exp(\kappa \tau)}, & \text{otherwise,} \end{cases} \tag{5.8}$$

where τ is a cosine threshold (e.g. $\tau = 0.9$) defining the cone aperture, and κ controls the rate of angular decay. This formulation softly encodes a conical field of view: points lying near the gaze direction retain high activation, while those further away decay exponentially.

The resulting 3D FoV field $V \in \mathbb{R}^{H \times W}$ takes the form of a 2D heatmap aligned with the original RGB image, since each pixel in the 2.5D point map corresponds to a unique spatial location in the image. Unlike 2D gaze cones, which highlight all pixels along the same image ray regardless of depth, this 3D representation enforces geometric consistency by weighting only points that are directionally aligned in 3D space as shown in Figure 5.1.

5.3.3 Scene Encoder

The Scene Encoder integrates the geometric and visual cues derived from the 3D field-of-view representation with the broader scene context. Its goal is to produce a person-conditioned representation of the scene that captures both appearance and geometry, enabling accurate localization of the gaze target and estimation of whether it lies inside or outside the image boundaries.

The input to the scene encoder consists of the RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the head box mask $\mathbf{M} \in \{0,1\}^{H \times W}$, and the 3D field-of-view map $\mathbf{V} \in \mathbb{R}^{H \times W}$ computed in the previous stage. These three signals are concatenated along the channel dimension and fed to a convolutional backbone $\mathscr{S}_b$, which extracts a *person-conditioned scene feature map*:

$$\mathbf{F}_{\text{scene}} = \mathscr{S}_b\big([\mathbf{I}, \mathbf{M}, \mathbf{V}]\big), \qquad \mathbf{F}_{\text{scene}} \in \mathbb{R}^{H' \times W' \times D}. \tag{5.9}$$

By conditioning the scene representation on the 3D FoV and head location, the backbone learns to emphasize spatial regions that are both visually salient and geometrically consistent

with the predicted gaze direction.

5.3.4 Gaze Decoder

The Gaze Decoder integrates information from the Gaze Encoder and Scene Encoder to produce the final gaze heatmap and in/out flag.

Gaze Heatmap Prediction. The feature map $\mathbf{F}_{\text{scene}}$ is passed to a convolutional decoder $\mathscr{S}_h$ that predicts the final gaze location heatmap $\hat{\mathbf{H}} \in \mathbb{R}^{H \times W}$:

$$\hat{\mathbf{H}} = \mathscr{S}_h(\mathbf{F}_{\text{scene}}). \tag{5.10}$$

This decoder refines the person-conditioned scene representation into a dense spatial distribution over potential gaze targets, combining the coarse geometric priors from the 3D FoV with fine-grained appearance cues from the image.

In/Out Prediction. The same scene feature map $\mathbf{F}_{\text{scene}}$ is further used to predict whether the gaze target lies within the image (*in*) or outside it (*out*). To this end, the feature map is first compressed into a 1D descriptor through an additional convolutional block followed by global max pooling:

$$\mathbf{x}_{\text{scene}} = \text{MaxPool}\big(\text{Conv}(\mathbf{F}_{\text{scene}})\big). \tag{5.11}$$

This vector encodes global contextual information about the scene in relation to the person. It is concatenated with the gaze embedding $\mathbf{x}_{\text{emb}}$ from the Gaze Encoder to form a joint representation:

$$\mathbf{x}_{\text{joint}} = [\mathbf{x}_{\text{scene}}, \mathbf{x}_{\text{emb}}], \tag{5.12}$$

which is passed through a MLP $\mathscr{S}_f$ to produce the in/out probability:

$$\hat{p}_{\text{in/out}} = \mathscr{S}_f(\mathbf{x}_{\text{joint}}). \tag{5.13}$$

This design ensures that the decision about whether the person is looking within or beyond the image is informed by both scene-level context and person-specific gaze cues.

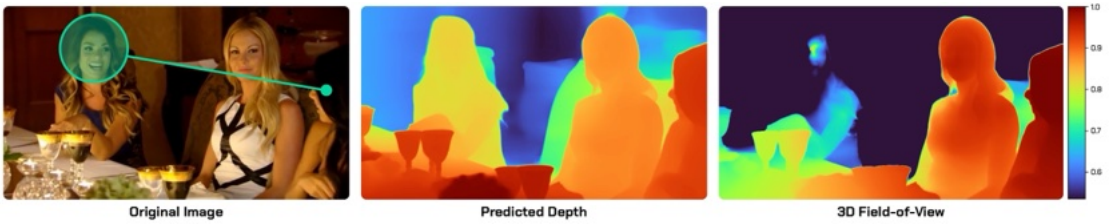


Figure 5.5: An example of a 3D field-of-view (FoV) derived from a ground-truth gaze point. **Left:** original image with the ground-truth head bounding box and gaze point. **Middle:** corresponding metric depth map. **Right:** resulting 3D FoV heatmap obtained after applying the exponential angular decay.

5.4 Implementation Details

Gaze Vector Supervision. Since gaze annotations are provided as 2D points in the image plane, they do not contain explicit 3D directional information. However, the 3D field-of-view representation introduced in the previous section requires supervision in the form of a 3D gaze vector expressed in the eye coordinate system $\mathbf{C}_e$. To obtain this, we derive a *pseudo ground-truth* 3D gaze vector from the annotated 2D gaze point using the estimated scene depth.

Given the 2D gaze annotation (u_g, v_g) and its corresponding depth value $D(u_g, v_g)$, we first back-project this pixel into the camera coordinate system $\mathbf{C}_c$ using Equation 5.6. The resulting 3D point $G_c \in \mathbb{R}^3$ corresponds to the annotated gaze target under the geometry of the reconstructed scene. To express this point in the same reference frame as the predicted gaze vector, we transform it to the local eye coordinate system $\mathbf{C}_e$ using Equation 5.5, producing $G_e \in \mathbb{R}^3$. Finally, the pseudo ground-truth 3D gaze vector $\mathbf{g}_v^{\text{gt}} \in \mathbb{R}^3$ is obtained by normalizing the direction from the origin of the eye frame to this target point:

$$\mathbf{g}_v^{\text{gt}} = \frac{G_e}{\|G_e\|}. \tag{5.14}$$

During training, it provides a geometric supervision signal for the Gaze Encoder, ensuring that the predicted direction $\mathbf{g}_v$ aligns with the physically plausible 3D ray passing through the annotated gaze point in the image. This approach yields a consistent 3D supervision signal even in the absence of true 3D gaze measurements. Although the derived vector relies on the estimated depth map and thus remains an approximation, it effectively constrains the network to learn a physically meaningful mapping between appearance cues and 3D gaze geometry. Refer to Figure 5.5 for an illustration of a 3D field-of-view derived from a ground-truth gaze point.

Loss Function. We optimize the model using a weighted sum of three objectives: (i) a pixel-wise MSE loss for gaze heatmap prediction, (ii) an angular loss for regressing the normalized gaze vector, and (iii) a binary cross-entropy loss for the in/out classification.

Model Configuration. The gaze backbone is a ResNet-18 (He et al., 2016) pre-trained on Gaze360 (Kellnhofer et al., 2019), which processes head crops resized to 224 × 224 pixels. The scene encoder employs an EfficientNet-B1 (Tan and Le, 2019) pre-trained on ImageNet (Deng et al., 2009), followed by a Feature Pyramid Network (Lin et al., 2017a), and operates on scene images of size 512 × 288. The heatmap decoder $\mathscr{S}_h$ comprises six convolutional layers and a final 1×1 regression layer, while the in/out prediction head $\mathscr{S}_f$ is implemented as a lightweight two-layer MLP.

Training Procedure. We train the model for 40 epochs on GazeFollow, and fine-tune it for an additional 20 epochs on VideoAttentionTarget. The same fine-tuning procedure is also applied to ChildPlay. Optimization is performed using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 2.5×10^{-4} during training on GazeFollow, and 2.5×10^{-5} during fine-tuning on VideoAttentionTarget and ChildPlay.

Inference Pipeline. During inference, the scene image is resized while preserving its original aspect ratio, with the longer side scaled to 512 pixels. The same depth estimation model used during training is applied to reconstruct the point cloud, ensuring geometric consistency with the pseudo 3D gaze supervision, which depends on the specific depth estimator employed.

5.5 Experimental Design

5.5.1 Benchmark Comparison

We report results on GazeFollow, VideoAttentionTarget, and ChildPlay in Table 5.1. Our model achieves competitive performance, on par with the state-of-the-art method of Jin et al. (2022), which also leverages depth information but without modeling an explicit 3D field-of-view. On GazeFollow, our approach attains a slightly higher average distance, matches their best performance on the minimum distance, and outperforms all prior works on AUC. On VideoAttentionTarget, both methods achieve identical distance error (0.108/0.109). Our method, however, yields noticeably lower average precision (AP) for the in/out classification task compared to previous works, suggesting that concatenating the early-stage gaze embedding with the late-stage person-conditioned saliency features may not be the most effective design choice.

On ChildPlay, we compare against our previous work Gupta et al. (2022), as the dataset was newly introduced in this study. The proposed 3D FoV formulation demonstrates stronger cross-dataset generalization, outperforming the prior model before and after fine-tuning. While fine-tuning improves both models, the earlier architecture shows a larger relative gain, suggesting mild overfitting to the dataset. In contrast, our method maintains more stable behavior and achieves the best performance across all metrics. Interestingly, the gap between model and human performance remains substantial, highlighting significant headroom for further progress.

Method	GazeFollow			VAT		ChildPlay	
	AUC ↑	Avg. Dist. ↓	Min. Dist. ↓	Dist. ↓	AP ↑	Dist. ↓	AP ↑
Recasens et al. (2015)	0.878	0.190	0.113	—	—	—	—
Chong et al. (2018)	0.896	0.187	0.112	0.171	0.712	—	—
Lian et al. (2018)	0.906	0.145	0.081	—	—	—	—
Chong et al. (2020b)	0.921	0.137	0.077	0.134	0.853	—	—
Fang et al. (2021)	0.922	0.124	0.067	**0.108**	0.896	—	—
Jin et al. (2021)	0.919	0.126	0.076	0.134	0.880	—	—
Jin et al. (2022)	0.920	**0.118**	**0.063**	0.109	**0.897**	—	—
Tonini et al. (2022)	0.927	0.141	—	0.129	—	—	—
Bao et al. (2022)	0.928	0.122	—	0.120	0.869	—	—
Hu et al. (2022b)	0.923	0.128	0.069	0.118	0.881	—	—
Gupta et al. (2022)†	0.933	0.134	0.071	0.137	—	0.142	—
Gupta et al. (2022)	0.933	0.134	0.071	0.122	0.864	0.113	0.983
GeomGaze†	**0.936**	0.125	0.064	0.123	—	0.115	—
GeomGaze	**0.936**	0.125	0.064	0.109	0.834	**0.107**	**0.986**
Human	0.924	0.096	0.040	0.051	0.925	0.048	0.993

Table 5.1: Results on the GazeFollow, VideoAttentionTarget, and ChildPlay datasets. The best scores are given in bold, while the second best scores are underlined. The † symbol means that the model was trained on GazeFollow and evaluated without fine-tuning.

It is also important to note that neither GazeFollow nor VideoAttentionTarget provides a dedicated validation set. Public implementations suggest that some prior works may have used test sets for model selection, potentially leading to optimistic results. In contrast, we explicitly carved out validation splits from the official training sets to ensure fair model development and reliable performance reporting.

5.5.2 Ablation Studies

To evaluate the contribution of the explicit 3D field-of-view (FoV) representation, we compare our approach against a variant that uses a standard 2D gaze cone across the three datasets. In this baseline, the 2D cone is obtained by computing the similarity between the projected 3D gaze vector and the 2D scene locations, without applying any angular decay.

The results in Table 5.2 show that *GeomGaze*, with its explicit 3D FoV formulation, performs on par or better than the 2D baseline across datasets. On GazeFollow, performance is slightly lower, likely because the dataset features relatively constrained scenes where depth variation is limited and the annotated person is almost always the one closest to the camera. In such settings, geometric depth cues bring little additional information. However, on VideoAttentionTarget and ChildPlay, the 3D FoV model demonstrates substantially stronger cross-dataset generalization and remains superior even after fine-tuning. These results underscore the importance of incorporating depth information and the advantage of a geometrically consistent 3D FoV representation for gaze target prediction.

Method	GazeFollow			VAT		ChildPlay	
	AUC ↑	Avg. Dist. ↓	Min. Dist. ↓	Dist. ↓	AP ↑	Dist. ↓	AP ↑
GeomGaze 2D FoV†	0.939	0.122	0.062	0.128	—	0.127	—
GeomGaze 2D FoV	0.939	0.122	0.062	0.120	0.856	0.121	0.944
GeomGaze 3D FoV†	0.936	0.125	0.064	0.123	—	0.115	—
GeomGaze 3D FoV	0.936	0.125	0.064	0.109	0.834	0.107	0.986

Table 5.2: Ablation experiment comparing 2D and 3D field-of-view (FoV) representations. The † symbol means that the model was trained on GazeFollow and evaluated without fine-tuning.

5.5.3 Stability of 3D Gaze Vector Supervision

We analyze the impact of depth quality on the reliability of 3D gaze vector supervision by comparing two monocular depth estimators: the model of Yin et al. (2021) and that of Patakin et al. (2022) used in our work. The goal is to assess whether the latter, which is trained to preserve scene geometry, produces more stable pseudo ground-truth gaze vectors.

The experiment relies on the assumption that a person's gaze vector, expressed in the camera coordinate system, should remain invariant to changes in image crop or apparent scale. For each selected frame, we take five random crops of the image and estimate depth and focal length independently for each crop. We then reconstruct the corresponding point clouds, derive the pseudo ground-truth 3D gaze vectors as described in Section 5.4, and compute the standard deviation of each gaze vector component across the five crops. To obtain robust estimates, this procedure is repeated for the first frame of every clip in the ChildPlay training set, and the median standard deviation is reported.

The results show that the method of Yin et al. (2021) yields median standard deviations of $[0.041, 0.032, 0.095]$, whereas the estimator of Patakin et al. (2022), which better preserves geometric consistency, achieves lower values of $[0.026, 0.019, 0.075]$. The improvement, particularly along the z-axis, indicates that using a geometry-preserving depth model leads to more stable and consistent 3D gaze vector supervision.

5.5.4 Qualitative Evaluation

Figure 5.6 presents qualitative predictions of *GeomGaze* on the *ChildPlay* dataset. The model accurately captures gaze direction in diverse, cluttered environments, whether toward another person or an object, demonstrating its ability to capture geometric scene information. The heatmaps remain spatially coherent and depth-aware, emphasizing regions consistent with the 3D gaze geometry. Interestingly, the predicted gaze points consistently align with high-intensity regions of the 3D field-of-view heatmap, indicating strong agreement between geometric reasoning and final localization.



Figure 5.6: Predictions of GeomGaze on images from the ChildPlay dataset. The person's head bounding box and predicted gaze point are shown in white, while the predicted 3D FoV heatmap is overlaid on the image. The predicted 3D gaze vector is projected onto the image and color-coded according to its depth orientation: blue when directed toward the foreground, and purple when oriented toward the background.

5.6 Limitations

While the proposed approach introduces an explicit 3D field-of-view formulation and demonstrates clear advantages over conventional 2D representations, several limitations remain.

Single-person formulation. The current framework operates in a single-person setting, where each gaze prediction is conditioned on one individual at a time. This formulation simplifies the problem and enables detailed geometric modeling, but it does not scale efficiently to crowded or multi-person scenes. At inference time, the scene must be re-encoded separately for each person, leading to redundant computation and higher latency as the number of people increases. The next chapter addresses this issue by introducing a novel design to handle multiple people efficiently within a shared scene representation.

Dependence on monocular depth and camera parameters. Our method relies on a monocular depth estimator and approximate camera intrinsics to reconstruct 3D scene geometry. Consequently, the quality of gaze supervision and 3D FoV construction is bounded by the accuracy of the predicted depth map and focal length. Although scale ambiguity is partially mitigated by using geometry-preserving depth models, absolute scale consistency remains imperfect. Moreover, as shown in our stability analysis, depth models differ in geometric fidelity, and using a different estimator at inference time may degrade performance unless the model is retrained or jointly adapted.

Implicit use of 3D FoV features. While the 3D field-of-view map injects geometric context into the scene encoder, it is concatenated as a feature channel rather than being explicitly modeled. This design allows flexibility but does not guarantee that the network learns to actively reason over visibility or spatial constraints. Future work could explore more principled fusion strategies or attention mechanisms that operate directly on 3D features.

Lack of occlusion reasoning. The 3D FoV representation highlights all points that are directionally aligned with the gaze vector, regardless of whether they are visible from the subject's point of view. A more physically grounded model would incorporate visibility reasoning (e.g. by performing ray-tracing over the 2.5D point map and stopping at the first intersection), to avoid attributing attention to occluded or unreachable regions. Exploring such formulations could lead to more faithful modeling of gaze geometry in complex scenes.

5.7 Conclusion

This chapter introduced the idea of reasoning about gaze through an explicit geometric representation of the scene. Rather than treating gaze as a purely image-based mapping, we reformulated the task in 3D by constructing a field-of-view representation that embeds the predicted gaze vector within the reconstructed scene geometry. This simple yet powerful concept brings depth awareness and spatial coherence to the problem, enabling the model to reason about gaze direction and potential targets in a physically meaningful way.

Although the proposed implementation relies on a conventional CNN backbone, the core contribution lies not in the architecture itself but in the underlying principle of geometric consistency. The 3D FoV formulation provides a modular and general mechanism that can be integrated into future designs to enrich them with explicit spatial reasoning. It highlights how even approximate 3D information derived from 2D annotation, when used coherently, can yield significant gains in robustness and interpretability.

At the same time, the current framework remains limited to single-person inference, which becomes computationally costly in multi-person scenes. The next chapter addresses this by rethinking the architectural design for efficiency and scalability, introducing a transformer-based model capable of jointly processing multiple people within a shared scene representation.

6 Efficiency & Multi-Person Scaling

Everyday social scenes are multi-person by default: classrooms, clinics, streets, homes, and workplaces rarely contain a single actor. Instead, they involve multiple people whose attention shifts rapidly across objects and one another. This chapter develops an architecture and toolkit for efficient, accurate, and scalable gaze following in such settings. It also marks a shift from conventional CNN-based designs to modern transformer-based architectures.

We begin by presenting the computational bottleneck at the heart of prior pipelines, situate existing approaches within related work, detail the proposed architecture (with an emphasis on amortized compute and multi-scale decoding), and evaluate it through a structured experimental design: benchmark comparisons, ablations, efficiency studies, cross-dataset transfer, and qualitative analyses of uncertainty and social cases (e.g. looking at humans, eye contact, shared attention). We conclude with limitations and a brief outlook to the next chapter, where we leverage the same principles to push from efficient multi-person inference toward richer semantic and social gaze understanding.

6.1 Motivation

Given the multi-person nature of relevant application domains, a practical gaze following system must therefore (i) handle several people simultaneously, (ii) scale without a linear explosion in computation, and (iii) remain accurate. Yet, most classic gaze following pipelines encode the entire scene repeatedly (i.e. once per person) because the input instance is defined as an (image, head) pair. This single-person formulation is convenient for modeling, but wasteful for inference: every additional person reruns the same heavy scene backbone, multiplies FLOPs and memory, and destroys latency budgets. It also weakens downstream integration, where gaze is only one module in a broader social cognition stack. Figure 6.1 shows the difference between two-stage single-person and two-stage multi-person frameworks.

This chapter addresses the efficiency bottleneck. To this end, we introduce Sharingan, a multi-person transformer architecture (Tafasca et al., 2024b) that encodes the scene once and

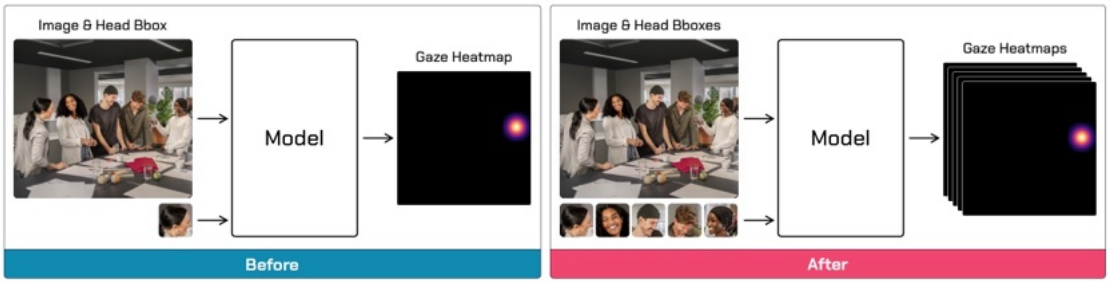


Figure 6.1: Illustration of the difference between previous methods that operate on (I, h) pairs, and our new method that takes all people as input $(I, h_1, h_2, .., h_n)$ and performs the expensive encoding once.

decodes per-person gaze outputs using compact, location-aware gaze tokens. This design preserves the original task formulation—predicting a heatmap and in/out flag for a specified person—while enabling true multi-person inference in a single forward pass. Compared to prior one-stage attempts that reformulate the problem as set prediction (and thus require post-hoc matching, complicating evaluation and integration), our approach stays compatible with existing benchmarks and plug-and-play within larger systems. Concretely, the architecture processes scene tokens and person-specific tokens together, and uses a lightweight, conditional multi-scale decoder to localize each person's gaze target. In effect, expensive compute becomes image-centric (amortized), and person-centric compute becomes light and parallelizable. This yields both speedups and better scaling with the number of people, without sacrificing accuracy on GazeFollow, VideoAttentionTarget, and ChildPlay, where it reaches or surpasses state of the art while remaining efficient at inference time.

The need for this shift is twofold:

- **Real-world throughput and latency.** Robotics and HRI demand reactivity; clinical and behavioral analytics demand throughput across long recordings. Recomputing the scene per person quickly becomes the dominant cost. Sharingan's shared scene encoding and per-person decoding markedly reduces redundant computation and memory traffic, enabling low-latency multi-person operation, even when adding modalities or temporal context later.
- **Social context is multi-person.** Social signals such as looking at humans, looking at each other (LAEO), and shared attention (SA) are intrinsically multi-agent. Architectures that natively represent people as tokens alongside scene tokens can model interactions and (in temporal variants) propagate constraints over time, which should improve both gaze localization and social inference.

Finally, we revisit the representation choice for gaze outputs. Direct coordinate regression is compact but collapses uncertainty and multi-modality into a single point; heatmap prediction naturally encodes ambiguity, precision–recall tradeoffs, and better supports semantic downstream tasks (e.g. verifying whether gaze lands on a head or object region). Within a multi-person transformer, a token-to-heatmap decoding pathway with conditional multi-scale fusion provides both localization fidelity and computational efficiency, outperforming heavier or mismatched decoders.

In summary, this chapter motivates and develops a compute-efficient, multi-person scaling strategy for gaze following that: (i) amortizes scene encoding across people; (ii) cleanly decodes per-person predictions at low marginal cost; (iii) maintains benchmark compatibility and accuracy; and (iv) forms a natural bridge toward temporal and social gaze modeling in subsequent chapters.

6.2 Related Work

Research on gaze following has traditionally focused on the single-person formulation, where each prediction instance is defined by a pair (I, h) of the full image and the head crop of the target person. Architectures typically adopt a dual-stream design: one branch encodes the scene to capture salient regions, while the other encodes the head to infer a coarse gaze direction, with a fusion module combining both into a final heatmap prediction. While effective, this approach has a fundamental drawback: the scene encoder must be re-executed for every person in the image, making it highly inefficient in multi-person scenarios.

6.2.1 Multi-Person Gaze Following

Several works have attempted to extend gaze following beyond the single-person setting. Early multi-person CNNs (e.g. Jin et al. (2021)) re-use a shared scene backbone fused with each head's features, but still process individuals independently, ignoring possible interactions. On the other hand, transformer-based set-prediction approaches such as Tu et al. (2022) and Tonini et al. (2023) predict head and gaze jointly from the scene image alone, inspired by DETR (Carion et al., 2020) as explained in Chapter 2. These methods natively handle multiple people, but introduce new challenges: they jointly predict head bounding boxes and gaze targets, thereby requiring a matching step to align predicted heads with annotations, complicating evaluation. Additionally, their formulation diverges from standard benchmarks, limiting direct comparison.

6.2.2 Efficiency and Decoding

In most gaze following architectures, the main design efforts have concentrated on the upstream components: how to encode person-specific gaze cues (e.g. from head crops or gaze cones), how to capture scene saliency (e.g. multimodal processing), and how to fuse these two streams. By contrast, the decoding stage, which is responsible for turning latent representations into the final gaze heatmap, has received far less attention. Typical approaches rely on a shallow stack of convolutional layers followed by transposed convolutions for upscaling (Chong et al. (2020b), Fang et al. (2021), Jin et al. (2022)). These modules are lightweight but constrained: they usually operate on very coarse spatial feature maps, often at 7×7, forcing the decoder to "hallucinate" fine-grained detail from downsampled activations. As a result, the predicted heatmaps often lack spatial precision, blur uncertainty, and fail to properly reflect

multimodality.

This limitation is particularly problematic in multi-person and social scenes, where subtle spatial cues (e.g. small head shifts, nearby competing targets) strongly affect interpretation. A coarse decoder bottleneck can obscure these distinctions, leading to overly diffuse predictions or mislocalized peaks. It also interacts with efficiency: if the decoder is weak, models are forced to rely on repeated, redundant computation upstream to recover detail. Designing stronger decoders that can leverage both scene and person tokens at multiple scales is therefore a crucial but underexplored lever for improving both accuracy and computational efficiency in gaze following.

6.3 Architecture

We now present the architecture developed for efficient multi-person gaze following. The design goal is to amortize costly scene encoding across all individuals, while maintaining explicit control over which people are included and producing accurate, person-specific predictions. Adopting a transformer backbone also raises the question of how to best encode the different input modalities: the global scene, the head crops, and their bounding boxes. To this end, the model represents each head as a gaze token that interacts with scene tokens within a transformer encoder. Predictions are then decoded using a conditional multi-scale decoder, yielding one gaze heatmap per person, along with a binary in/out label. We provide an overview of the architecture in Figure 6.2.

6.3.1 Scene Tokens

The first input to the model is the scene image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, which is decomposed into non-overlapping patches and projected into a sequence of scene tokens $\mathbf{x}_s \in \mathbb{R}^{N \times D}$, where N is the number of patches and D is the token dimension. Each token is enriched with positional information, allowing the model to maintain spatial awareness of the scene layout. These image tokens serve as a global representation of the environment, capturing contextual information such as salient objects, spatial structure, and other people.

6.3.2 Gaze Tokens

Each annotated head in the image is encoded as a compact gaze token. To construct it, the head crop $\mathbf{h}_{\text{crop}} \in \mathbb{R}^{h \times w \times C}$ is passed through a gaze backbone to obtain a person-specific gaze embedding $\mathbf{g}_{\text{emb}} \in \mathbb{R}^{d_{\text{emb}}}$. This embedding is processed by a gaze regressor to produce a 2D gaze vector $\mathbf{g}_v \in \mathbb{R}^2$ which we supervise using an angular loss. At the same time, the gaze embedding $\mathbf{g}_{\text{emb}}$ is also projected to the token dimension, leading to $\mathbf{x}_{\text{emb}} \in \mathbb{R}^D$. In addition, the head bounding box coordinates $\mathbf{h}_{\text{bbox}} \in [0,1]^4$ are linearly projected to $\mathbf{x}_{\text{bbox}} \in \mathbb{R}^D$ in order to encode the spatial location and scale of the person within the image. The sum of

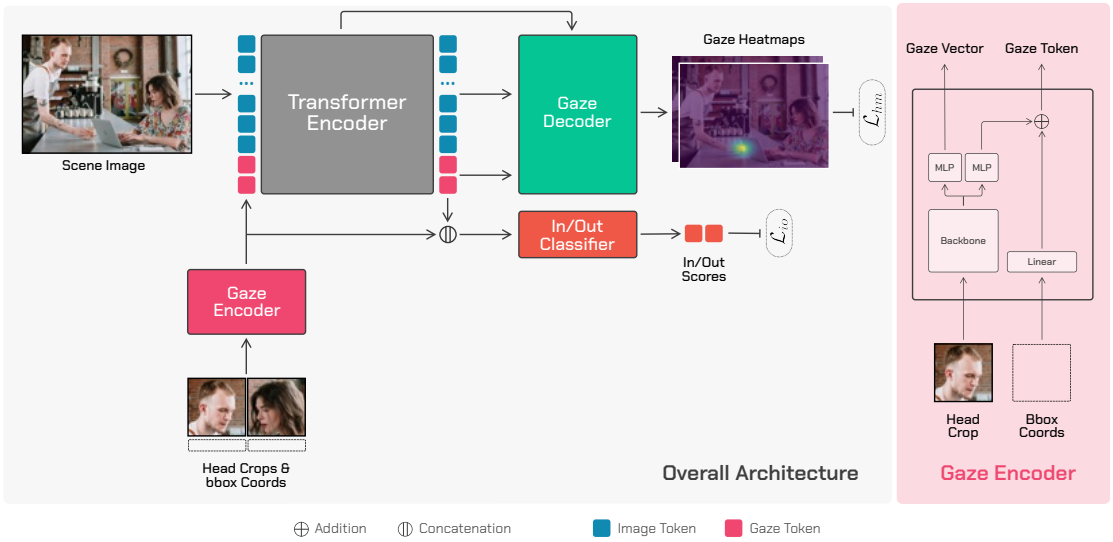


Figure 6.2: Overview of our Sharingan architecture. **A.** The input scene is divided into patches and projected into image tokens. **B.** Each head crop and its bounding box are processed to create a location- and size-aware gaze token. Specifically, the head crop passes through a gaze backbone to produce an embedding that is (i) used to predict a normalized 2D gaze vector supervised with an angular loss, and (ii) projected to the token dimension. The head bounding box is also projected into an embedding, which is added to the gaze token to encode spatial location. **C.** Scene and gaze tokens are concatenated and processed jointly by the transformer encoder. The resulting tokens corresponding to each person are then decoded with a gaze decoder (Conditional DPT) to produce per-person gaze heatmaps. Finally, the input and output gaze tokens are concatenated and fed to a classifier to predict whether the person is looking inside or outside the frame.

both embeddings forms a location-aware gaze token: $\mathbf{x}_{\text{g}} = \mathbf{x}_{\text{emb}} + \mathbf{x}_{\text{bbox}} \in \mathbb{R}^D$. This gaze token captures appearance cues (e.g. head pose, eyes, orientation) and gaze direction information via auxiliary supervision.

Multi-Person Setting. When $N_p > 1$ people are used, the architecture produces a set of gaze tokens independently (one per individual). To simplify notation, we use $\mathbf{x}_{\text{g}} = \mathbf{x}_{\text{g},1} \oplus \mathbf{x}_{\text{g},2} \ldots \oplus \mathbf{x}_{\text{g},N_p} \in \mathbb{R}^{N_p \times D}$ to also denote the concatenation of all gaze tokens.

Modality Encoding. Since gaze tokens and scene tokens originate from different sources, the model must be able to distinguish between the two modalities. Instead of introducing an explicit modality encoding, we rely on the projection layers themselves: the learnable bias terms of the gaze and scene projections naturally encode modality-specific information, allowing the transformer to differentiate inputs without additional mechanisms.

This design provides the model with explicit control over which people are considered, and stands in contrast to DETR-style formulations that predict heads implicitly and require post-hoc matching. Here, gaze tokens are fully supervised and aligned with the annotated people.

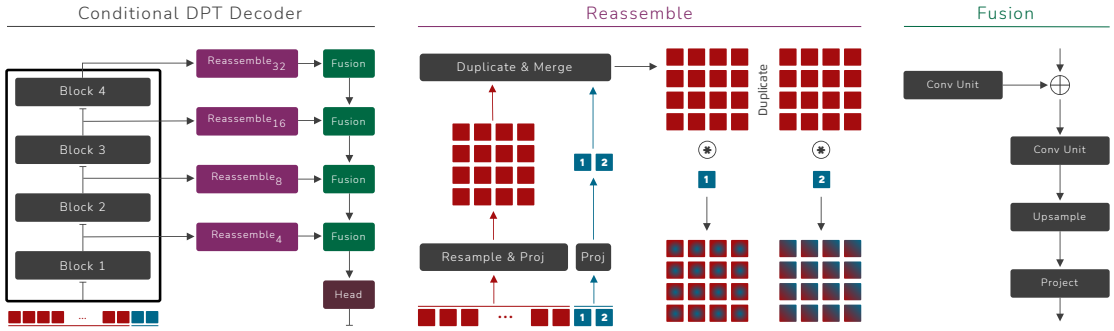


Figure 6.3: Overview of the Conditional DPT decoder. **Left:** Four intermediate outputs from the transformer encoder are reassembled into image-like feature maps at resolutions $(\frac{H}{k} \times \frac{W}{k})$ and fused progressively. **Middle:** In the Reassemble module, the scene feature map is conditioned on gaze tokens: features are duplicated for all people, combined with projected gaze embeddings through element-wise multiplication, and merged into person-specific feature maps. **Right:** The fusion module integrates features across scales using residual convolutional units, upsampling, and projection layers, ultimately producing one gaze heatmap per person. This design allows predictions to leverage both coarse global context and fine spatial detail, while remaining explicitly conditioned on individual gaze tokens.

6.3.3 Scene Encoder

The transformer encoder follows the standard ViT design (Dosovitskiy et al., 2020). Its input is the concatenation of scene tokens and the set of gaze tokens $\mathbf{x} = \mathbf{x}_s \oplus \mathbf{x}_g \in \mathbb{R}^{N_t \times D}$, where $N_t = N + N_p$. This combined sequence is then processed by a stack of L self-attention-based transformer blocks (Vaswani et al., 2017), producing an output of the same dimensionality: $\mathbf{x}^{(L)} \in \mathbb{R}^{N_t \times D}$.

By embedding people and scene into a shared token space, the encoder allows gaze tokens to attend to relevant regions of the scene and scene tokens to incorporate person-specific cues, enabling joint reasoning about where each individual is likely to be looking. Crucially, this step is performed only once per image, regardless of the number of people, which ensures computational efficiency.

6.3.4 Gaze Decoder

The role of the gaze decoder is to generate a set of gaze heatmaps, one per person. To achieve this, we introduce a Conditional DPT decoder (*cf.* Figure 6.3) that progressively fuses intermediate representations of scene tokens $\mathbf{x}_s^{(i)}$ and gaze tokens $\mathbf{x}_g^{(i)}$ at multiple *simulated* resolutions. Lower resolutions have higher channel capacity and correspond to deeper layers of the encoder, making the design analogous to an isotropic variant of a Feature Pyramid Network (Lin et al., 2017b).

Our approach is inspired by DPT (Ranftl et al., 2021), which is designed to decode only image tokens. In contrast, we require predictions to be explicitly conditioned on each person. To this end, after each selected encoder block $l_k \in \{4, 8, 16, 32\}$, the scene tokens $\mathbf{x}_s^{(l_k)}$ are reassembled into an image-like feature map of resolution $(\frac{H}{k}, \frac{W}{k})$ and channel dimension d_k

The corresponding gaze tokens $\mathbf{x}_g^{(l_k)}$ are projected into the same dimensionality.

For each person, the gaze token is combined with the reassembled image map via an element-wise multiplication. Since this operation must be repeated for all individuals, we first duplicate the image feature maps N_p times, then compute the token–map interaction for each copy. The resulting person-specific feature maps are stacked and merged, yielding a tensor of shape $(B \times N_p, d_k, \frac{H}{k}, \frac{W}{k})$.

This tensor is passed to a fusion module consisting of a residual convolutional block, which integrates information from the previous fusion stage. The output is then processed by another residual block, followed by an upsampling step that doubles the resolution and a projection layer that adjusts the channel dimension. At stage k, this produces a tensor of shape $(B \times N_p, d_{\frac{k}{2}}, \frac{2H}{k}, \frac{2W}{k})$.

The process continues through progressively finer resolutions, ultimately yielding a feature tensor of size $(B \times N_p, d_{\text{out}}, \frac{H}{2}, \frac{W}{2})$. A final convolutional head reduces the channel dimension to 1 and resizes the spatial dimensions to match the target heatmap size. After separating the batch and person dimensions, the output takes the form $(B, N_p, 1, H_{\text{hm}}, W_{\text{hm}})$, i.e. one heatmap per person.

The rationale for this design is twofold. First, by leveraging multi-scale features, the decoder preserves both coarse contextual information and fine spatial detail, which is an essential property for dense prediction tasks. Second, conditioning image features on gaze tokens ensures that predictions are explicitly person-specific, allowing the model to handle multiple individuals in a single forward pass. Notably, early-layer features often retain more modality-specific information (e.g. scene-independent gaze cues like head orientation for gaze tokens) that remain useful when fused with later, context-rich layers, leading to more accurate and robust heatmap predictions.

6.3.5 In/Out Decoder

The In/Out classification module is implemented as a 7-layer MLP. It takes as input the concatenation of the initial and final gaze tokens $\mathbf{x}_g^{(0)} \oplus \mathbf{x}_g^{(L)}$, and outputs a binary label indicating whether each person is looking inside or outside the image frame.

6.4 Implementation Details

Loss Function. We optimize the model using a weighted sum of three objectives: (i) a pixel-wise MSE loss for gaze heatmap prediction, (ii) an angular loss for regressing the normalized gaze vector, and (iii) a binary cross-entropy loss for the in/out classification.

Model Configuration. Inputs consist of the full scene image and head crops, both resized to 224×224. The output gaze heatmaps have a resolution of 64×64. The gaze backbone is

a ResNet-18 (He et al., 2016) pretrained on Gaze360 (Kellnhofer et al., 2019), while the scene encoder is a ViT-Base model (Dosovitskiy et al., 2020) initialized with weights from MultiMAE (Bachmann et al., 2022).

Context People. Training requires the model to process multiple people jointly. To make this tractable, we fix the number of persons per training batch to N_p^{train}. For each image, we always include the ground-truth annotated person, and when available, we randomly sample $N_p^{\text{train}} - 1$ additional detected heads. If fewer heads are present, we pad with empty boxes and crops. Importantly, the loss is only computed for the annotated person.

At evaluation, the setup is more flexible: for each image k, the number of processed people $N_{p,k}^{\text{eval}}$ is set to the total number of heads detected in that frame. All are handled simultaneously in a single forward pass (batch size 1). Note that N_p^{train} is only a training constraint and does not limit the number of people at inference. Unless stated otherwise, we use $N_p^{\text{train}} = 2$.

Training Procedure. We train the model for 20 epochs on GazeFollow. For VideoAttention-Target and ChildPlay, we follow standard practice: the GazeFollow-trained model is used for initialization, after which we freeze all parameters except the gaze decoder and in/out decoder, which are fine-tuned for 2 epochs each. Training uses AdamW with a base learning rate of 3×10^{-5} and a cosine annealing schedule. During fine-tuning, the learning rate is set to 1×10^{-6} for the gaze decoder and 3×10^{-4} for the in/out classifier.

6.5 Experimental Design

6.5.1 Benchmark Comparison

We report results on GazeFollow, VideoAttentionTarget and ChildPlay in Table 6.1, comparing against prior work. Sharingan achieves state-of-the-art performance across all three datasets on most metrics, and improves over the only comparable multi-person baseline of Jin et al. (2021) by 0.013 on Avg. Dist. and 0.020 on Min. Dist.

Our model achieves strong performance even compared to multimodal methods. For example, the best single-person competitor is the multimodal approach of Bao et al. (2022), which relies on additional modalities (depth and pose), making it even more expensive during inference given the restriction to single-person. We outperform their method on both GazeFollow (by 7% on Avg. Dist.) and VideoAttentionTarget (by 11% on Dist.).

On the video benchmarks (VideoAttentionTarget and ChildPlay), Sharingan continues to perform strongly, notably when trained on GazeFollow alone: without fine-tuning, it already surpasses multi-person competitors by a wide margin. Our previous geometric method GeomGaze (Tafasca et al., 2023b) performs on par with Sharingan on these datasets, likely because the depth-based geometric prior benefits from higher image quality, though that approach remains multimodal and single-person.

Method	GazeFollow			VAT		ChildPlay	
	AUC ↑	Avg. Dist. ↓	Min. Dist. ↓	Dist. ↓	AP ↑	Dist. ↓	AP ↑
Recasens et al. (2015)	0.878	0.190	0.113	—	—	—	—
Chong et al. (2018)	0.896	0.187	0.112	0.171	0.712	—	—
Lian et al. (2018)	0.906	0.145	0.081	—	—	—	—
Chong et al. (2020b)	0.921	0.137	0.077	0.134	0.853	—	—
Fang et al. (2021)	0.922	0.124	0.067	0.108	0.896	—	—
Jin et al. (2021)$_{\mp}$	0.919	0.126	0.076	0.134	0.880	—	—
Jin et al. (2022)	0.909	0.137	0.077	—	—	—	—
Tonini et al. (2022)	0.927	0.141	—	0.129	—	—	—
Bao et al. (2022)	0.928	0.122	—	0.120	0.869	—	—
Hu et al. (2022b)	0.923	0.128	0.069	0.118	0.881	—	—
Gupta et al. (2022)	0.933	0.134	0.071	0.122	0.864	0.113	0.983
GeomGaze	0.936	0.125	0.064	0.109	0.834	0.107	0.986
Sharingan$^{\dagger}_{\mp}$	**0.944**	**0.113**	**0.057**	0.113	—	0.109	—
Sharingan$_{\mp}$	**0.944**	**0.113**	**0.057**	**0.107**	**0.891**	**0.106**	**0.990**
Human	0.924	0.096	0.040	0.051	0.925	0.048	0.993

Table 6.1: Results on the GazeFollow, VideoAttentionTarget (VAT), and ChildPlay datasets. The best scores for multi-person models are given in bold, while the best scores in general are underlined. The $\mp$ symbol means the model is multi-person, while the † symbol means that the model was trained on GazeFollow and evaluated without fine-tuning.

6.5.2 Ablation Studies

Person Encoding. A core design choice in Sharingan is how people are represented within the architecture. Earlier works often encode head location as a binary mask, and gaze information as either a visual attention map (Chong et al., 2020b) or a gaze cone (Gupta et al., 2022; Fang et al., 2021; Tafasca et al., 2023b).

To analyze this component, we modify Sharingan and test three alternative person encoding strategies. We opt for single-person setting to keep the setup simple. The first baseline, *Head Mask Embed*, uses the box coordinates to construct a binary head mask, tokenize it, and add the resulting embeddings to the image tokens. The second baseline, *Head Crop Tokens*, keeps the same head mask embedding for spatial location, and tokenizes the head crop itself before appending directly to the image tokens. The third baseline, *Gaze Cone Tokens*, avoids processing the head directly, and instead, relies on the gaze backbone to regress a gaze vector, construct a corresponding 2D gaze cone, tokenize it, and append to the image tokens.

Note that the second and third variants approximately double the number of tokens compared to the baseline. The outputs of all variants are decoded into gaze heatmaps using a standard DPT decoder (Ranftl et al., 2021). As shown in Table 6.4a, our proposed location-aware gaze token is both the most efficient and achieves the best performance, confirming its suitability as the default person representation.

Gaze Decoder. We also evaluate the impact of the decoding mechanism by comparing our

Method	AUC ↑	Avg. D. ↓	Min. D. ↓
Head Mask Embed	0.940	0.117	0.060
Head Crop Tokens	0.933	0.138	0.076
Gaze Cone Tokens	0.934	0.133	0.073
Ours (Gaze Token)	**0.944**	**0.113**	**0.057**

(a) Person Encoding.

Method	AUC ↑	Avg. D. ↓	Min. D. ↓
Token to Heatmap	0.647	0.302	0.234
Dot-Product	0.923	0.120	0.062
Up & Dot-Product	0.934	0.116	0.059
Ours (Cond. DPT)	**0.944**	**0.113**	**0.057**

(b) Gaze Decoding.

Figure 6.4: Ablation results for core architectural design choices.

Conditional DPT decoder against simpler alternatives (Table 6.4b). The first baseline, *Token to Heatmap*, regresses the gaze heatmap directly from each output person token via an MLP, similar to the decoding strategies used in Tu et al. (2022) and Tonini et al. (2023). The second baseline, *Dot-Product*, projects both person and image tokens, then computes their element-wise dot-product before resizing the result to 64×64. The third baseline, *Up & Dot-Product*, upsamples the image representation first and then applies the dot-product operation.

Across all comparisons, our Conditional DPT decoder consistently outperforms the alternatives. This highlights the importance of using a decoder capable of leveraging multi-scale information and conditioning predictions on both person and scene tokens, rather than relying solely on direct regression or simple dot-product interactions.

6.5.3 Efficiency Analysis

A key advantage of our architecture is its ability to handle multiple people jointly, thereby amortizing the heavy scene encoding across all individuals. This property translates into improved computational efficiency as the number of people increases.

Figure 6.5 compares the FLOPs of our method with two representative baselines: Chong et al. (2020b), a lightweight two-stream CNN architecture, and Tonini et al. (2023), a one-stage DETR-style architecture. For fairness, the FLOPs count for our method and Chong et al. (2020b) includes the head detection step.

Despite being a one-stage model, Tonini et al. (2023) has a relatively high and nearly constant FLOPs count. This can be explained by two reasons: (i) the DETR-style design, where fixed learnable queries must be processed against the entire image representation at every forward pass, regardless of the number of people actually present, and (ii) the high resolution of the input image coupled with the self-attention mechanism that scales quadratically with the number of tokens. In contrast, Chong et al. (2020b) processes each person independently through its two-stream design, leading to a FLOPs curve that grows linearly with N_p^{eval} but starts from a relatively low base.

Our approach strikes a favorable balance: since the scene is encoded only once, the FLOPs increase with N_p^{eval} comes mainly from the lightweight gaze encoder/decoder and from the increased token count in the transformer encoder. This results in a curve that increases more gracefully with N_p^{eval} than Chong et al. (2020b), while remaining significantly lower than Tonini

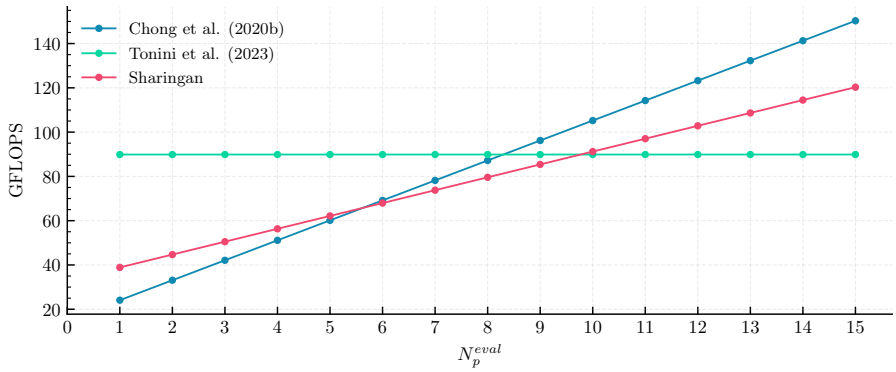


Figure 6.5: Comparative analysis of FLOPS vs number of people during inference.

et al. (2023) for a wide range of values.

Finally, we note that the ViT-Base backbone used in our model has 86M parameters, which is considerably larger than the ResNet-50 backbones (25M parameters) employed in both Chong et al. (2020b) and Tonini et al. (2023). This means that our FLOPs comparison is conservative: if backbone sizes were controlled to be equivalent, the relative efficiency gains of Sharingan would be even more pronounced.

6.5.4 Role of Training Group Size

Another distinctive property of our architecture is that it imposes *no hard limit* on the number of people that can be processed at inference. The model is permutation-equivariant with respect to the set of gaze tokens: each person is represented by a symmetric token, and the transformer encoder naturally accommodates any number of such tokens in the input sequence. Thus, at test time we can simply set N_p^{eval} to the actual number of detected people in an image, and process them jointly in a single forward pass.

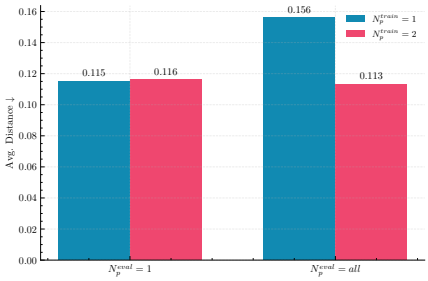
During training, however, we must fix the number of people per batch to a constant N_p^{train} for computational reasons (e.g. GPU memory). It is important to stress that this is *only* a batching constraint: the architecture itself remains agnostic to the choice of N_p^{train}.

Single-person vs. multi-person training. The fundamental distinction arises between training with a single person ($N_p^{\text{train}} = 1$) and with more than one person ($N_p^{\text{train}} > 1$). When trained in the single-person regime, the model inevitably learns to *anchor* its scene representation to a single gaze token. Because it never experiences multiple tokens in parallel, it has no incentive to learn how to distribute scene features across gaze tokens. As a result, when evaluated with $N_p^{\text{eval}} > 1$, its performance degrades significantly.

In contrast, training with at least two people ($N_p^{\text{train}} \geq 2$) introduces the inductive bias of *"more than one"*. The model must now share image features across multiple tokens and resolve them symmetrically, since gaze tokens are interchangeable and carry no person order encoding. Once this mechanism is learned, it generalizes seamlessly to different values of

	AUC ↑	Avg. D. ↓	Min. D. ↓
$N_p^{\text{train}} = 2$	0.944	0.113	0.057
$N_p^{\text{train}} = 3$	0.943	0.114	0.058
$N_p^{\text{train}} = 4$	0.942	0.113	0.057

(a) Ablation of N_p^{train} (eval: N_p^{eval} = all).



(b) Single- vs multi-person training.

Figure 6.6: Effect of the number of people during training on GazeFollow.

N_p^{eval}: the same machinery that works for two tokens also works for three, four, or more. It is worth emphasizing that even when $N_p^{\text{train}} > 1$, supervision is applied only to the annotated person in each image; the additional people serve purely as contextual tokens, and although unsupervised, their presence is essential to expose the model to multi-person inputs during training.

Figure 6.6b compares a single-person model ($N_p^{\text{train}} = 1$) and a multi-person model ($N_p^{\text{train}} = 2$) under two evaluation protocols: using only one person per image ($N_p^{\text{eval}} = 1$) and using all available people ($N_p^{\text{eval}} = \text{all}$). The multi-person model achieves consistent performance in both settings, while the single-person model performs well only when evaluated with a single person, but suffers a clear degradation when asked to handle multiple people. This confirms that training with $N_p^{\text{train}} > 1$ is crucial for generalization to realistic multi-person scenarios.

How many people are needed during training? An ablation study varying N_p^{train} from 2 to 4 (Table 6.6a) shows that performance remains essentially unchanged. This indicates that the *qualitative leap* occurs at $N_p^{\text{train}} = 2$: once the model has learned to process multiple gaze tokens, adding more during training yields no further benefit, only additional computational cost.

Practical considerations. Although the architecture can in principle scale to arbitrary N_p^{eval}, in practice extremely large numbers of people may introduce limitations due to GPU memory, diluted attention across gaze tokens, or other artifacts. Nevertheless, for realistic values of N_p^{eval} encountered in gaze following benchmarks (typically up to 5–10 people), a model trained with $N_p^{\text{train}} = 2$ is sufficient and generalizes effectively.

6.5.5 Output Representation

Gaze following is most commonly formulated as a heatmap prediction task. To the best of our knowledge, Lian et al. (2018) are the only authors to briefly consider a coordinate regression objective in their ablation study. Here, we revisit this formulation to better understand its properties and limitations.

Heatmap vs. Coordinate Regression. We train a variant of Sharingan where the Conditional

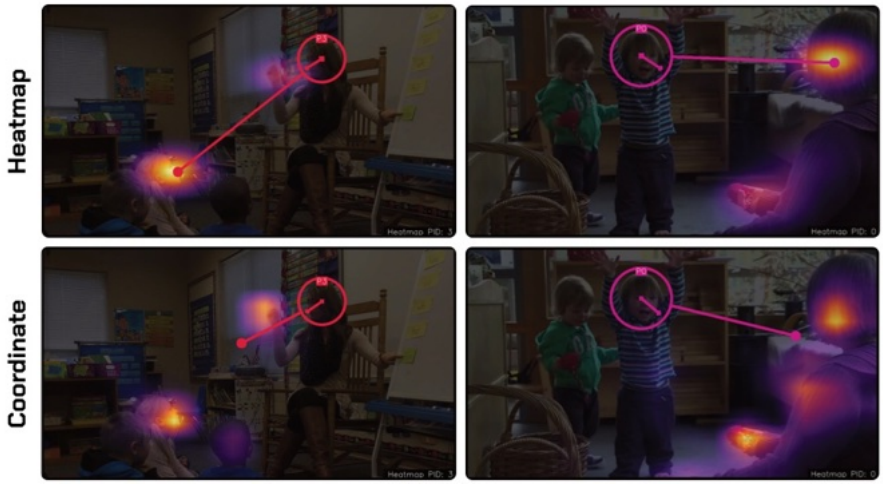


Figure 6.7: Comparison between heatmap and coordinate regression models. For the coordinate regression model, we use the attention map from the last layer of the transformer encoder as a proxy for the output heatmap. For the heatmap model, the gaze point corresponds to the argmax. We show the head bounding box, the predicted gaze point, gaze vector, and gaze/attention heatmap.

DPT decoder is replaced by a simple MLP that directly regresses (x, y) coordinates from the output gaze token of each subject. Unlike a heatmap, this coordinate-based model produces only a single point and therefore cannot represent the full probability distribution over the 2D image plane. Instead, it likely converges to the *expectation* of the posterior distribution $p(y \mid x)$, where y denotes the gaze point and x the input. In cases where the distribution is multimodal (i.e. multiple plausible gaze targets), the expectation may correspond to a location that is not itself likely under the distribution, but rather a weighted average of the modes.

This effect is illustrated in Figure 6.7. Since the coordinate model does not output a heatmap, we visualize the last-layer attention weights from the person's gaze token to the image tokens. This is not a true gaze distribution, but can serve as a proxy to highlight the regions the model attends to. It is clear that both the heatmap and coordinate regression models capture the main modes of the distribution in the two images. However, while the heatmap preserves multimodality explicitly, the coordinate regression model collapses these modes into a single prediction, which often falls between them and away from any true target.

Coordinate Regression as Posterior Expectation. Let $p(y \mid x)$ denote the true conditional distribution of gaze locations $y \in \mathbb{R}^2$ given a model input x (image and person-specific information). A coordinate regression model parameterized by θ outputs a prediction $\hat{y}(x;\theta) \in \mathbb{R}^2$. The model is trained by minimizing the mean squared error (MSE) between the predicted coordinate $\hat{y}(x;\theta)$ and ground-truth samples $y \sim p(y \mid x)$:

$$\mathscr{L}(\theta \mid x) = \mathbb{E}_{y \sim p(y|x)} \left[\| y - \hat{y}(x;\theta) \|^2 \right].$$

Expanding the expectation yields:

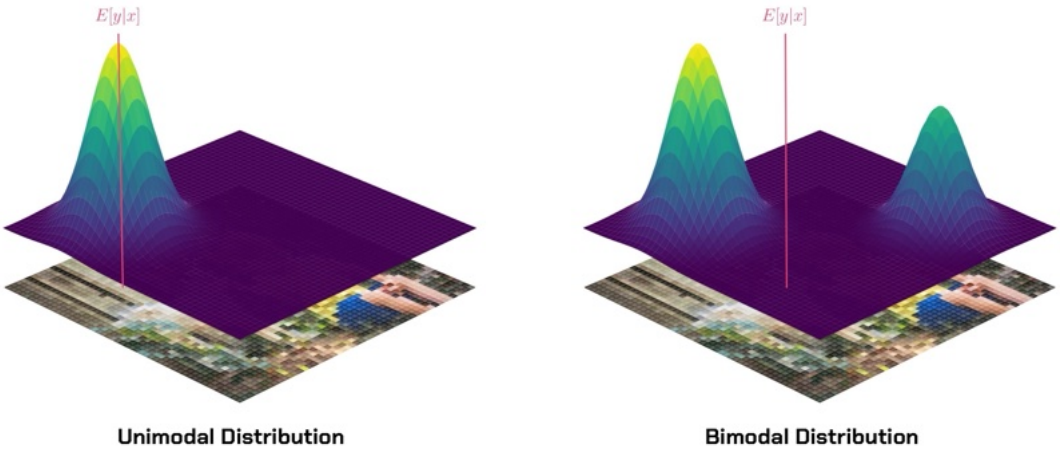


Figure 6.8: Unimodal vs. bimodal distributions. On the left, the unimodal case: the conditional expectation coincides with the mode, yielding a plausible prediction. On the right, the bimodal case: the conditional expectation lies between the two peaks, outside either mode, and thus corresponds to an unlikely target. This illustrates why direct regression to the mean can produce implausible estimates when the underlying distribution is multimodal.

$$\mathscr{L}(\theta \mid x) = \mathbb{E}[\|y\|^2] - 2\,\hat{y}(x;\theta)^\top \mathbb{E}[y] + \|\hat{y}(x;\theta)\|^2.$$

Taking the gradient with respect to the prediction $\hat{y}(x;\theta)$ and setting it to zero gives:

$$\nabla_{\hat{y}} \mathscr{L}(\theta \mid x) = -2\,\mathbb{E}[y] + 2\,\hat{y}(x;\theta) = 0,$$

which leads to the solution:

$$\hat{y}(x;\theta) = \mathbb{E}[y \mid x].$$

Therefore, a coordinate regression model trained with MSE learns the *conditional expectation* of the gaze distribution. When $p(y \mid x)$ is unimodal, this expectation coincides with the most likely gaze location. However, if the distribution is multimodal, which is often the case in gaze following, the expectation may lie between modes, resulting in predictions that are unlikely under $p(y \mid x)$, which is precisely the behavior observed in practice. We illustrate this behavior in Figure 6.8.

This effect is especially pronounced in gaze following, where multimodality is intrinsic and naturally arises whenever several targets are plausible. By contrast, other keypoint detection tasks such as human pose estimation are typically unimodal, since a body joint usually has a single, visually distinctive, well-defined location. In those cases, multimodality appears mainly due to limited training data, poor image quality, or occlusions. In short, multimodality

	Avg. Dist. ↓	Min. Dist. ↓	P_{LAH} ↑	R_{LAH} ↑
Coordinate Regression	0.106	0.066	0.683	0.368
Heatmap Regression	0.113	0.057	0.667	0.571

Table 6.2: Comparison between the *heatmap* and the *coordinate* training objectives on GazeFollow. P_{LAH} and R_{LAH} are the precision and recall of LAH, respectively.

is the *norm* in gaze following but the *exception* in most other keypoint-based tasks.

Quantitative Evaluation. The results in Table 6.2 further confirm this behavior. The coordinate regression model achieves better *average distance* and P_{LAH} scores, since these metrics are aligned with predicting a central tendency across multiple annotations. However, it lags behind the heatmap model in terms of *minimum distance* and R_{LAH}. This is expected: when the ground-truth annotations are diverse, the average distance reflects a mean-like target that favors the coordinate regression model. At the same time, positive LAH predictions from the coordinate model are highly confident (i.e. monomodal distribution), leading to improved P_{LAH}, but this also results in more false negatives. As a consequence, its R_{LAH} drops significantly, since predictions often fall just outside the annotated region when uncertainty is high.

Taken together, these results suggest that the *average distance* metric alone is not sufficient for evaluating gaze following models, a claim also supported by Bao et al. (2022). While coordinate regression may perform well on central-tendency metrics, heatmap prediction remains the more faithful formulation for capturing the inherent multimodality of gaze.

6.5.6 Generalization and Applications

To further evaluate the robustness of Sharingan, we test it on additional gaze-related tasks beyond standard gaze following benchmarks.

Gaze Following. On DLGaze, which differs substantially from GazeFollow in both setting and appearance (*cf.* Chapter 3), Sharingan (trained only on GazeFollow) achieves competitive distance scores (Table 6.9a), confirming strong generalization to unseen contexts.

Shared Attention. On VideoCoAtt, we adapt Sharingan's predictions to detect and localize shared attention. Results in Table 6.9b show clear improvements over prior work, with stronger F-Score@1.6, AP, and distance metrics, demonstrating that a multi-person gaze following model can be readily extended to higher-level social attention.

Mutual Gaze. On UCO-LAEO, Sharingan's predictions are used to infer mutual gaze relations between pairs of people by post-processing the gaze points. As shown in Table 6.9c, our method significantly outperforms baselines in F-Score, highlighting its ability to capture direct interpersonal gaze interactions.

Method	Dist. ↓
Random	0.442
Chong et al. (2020b)	0.138
Sharingan	**0.124**

(a) DL Gaze.

Method	F-score@1.6 ↑	AP ↑	Dist. ↓
Random	—	—	186
Chong et al. (2020b)	29.14	36.35	68
Sharingan	**46.19**	**42.96**	**55**

(b) VideoCoAtt.

Method	F-score ↑
Random	47.74
Chong et al. (2020b)	79.84
Sharingan	**84.79**

(c) UCO-LAEO.

Figure 6.9: Generalization experiments on DL Gaze (gaze following), VideoCoAtt (shared attention), and UCO-LAEO (mutual gaze).

These results illustrate that our method not only advances gaze following but also transfers naturally to downstream tasks of social gaze analysis, paving the way for broader applications in understanding group behavior.

6.5.7 Qualitative Evaluation

Figure 6.10 shows representative predictions of Sharingan across GazeFollow, VideoAttentionTarget, and ChildPlay, along with typical failure cases. The second column illustrates that the model can jointly predict the gaze points of all people in the scene, often capturing subtle cases such as multiple individuals converging on the same object (rows 2, 4 and 5) or diverging across the scene (row 3). The third column highlights the predicted heatmap for a selected person, which is generally sharp and well localized on plausible gaze targets. The heatmaps also appear well calibrated, with their intensity reflecting the likelihood of a region being the true gaze target (rows 3 and 6). In general, we observe strong generalization to diverse scenes, activities, objects and people. Failure cases arise either from dataset biases—such as GazeFollow's tendency to over-annotate hands, causing the model to snap to nearby hands (rows 7 and 8)—or from scenarios requiring complex reasoning, for instance when a person looks at a distant object indicated by another's pointing gesture (row 8).

6.6 Limitations

Sharingan processes scene tokens and gaze tokens jointly in the transformer encoder. While effective, this design means the same weights operate on both modalities, making it difficult to interpret how person-specific and scene-specific information are being combined. A more structured alternative would be to disentangle person and scene encoding into separate streams and fuse them selectively across the architecture. Such a design could improve interpretability, enhance stability in video settings, and enable richer modeling of person–person interactions. It would also address an efficiency limitation of Sharingan: in the current formulation, adding more people to the image increases the number of gaze tokens passed to self-attention, which scales quadratically with sequence length. Disentangling person processing could therefore reduce this overhead and make the model scale more gracefully to crowded scenes.

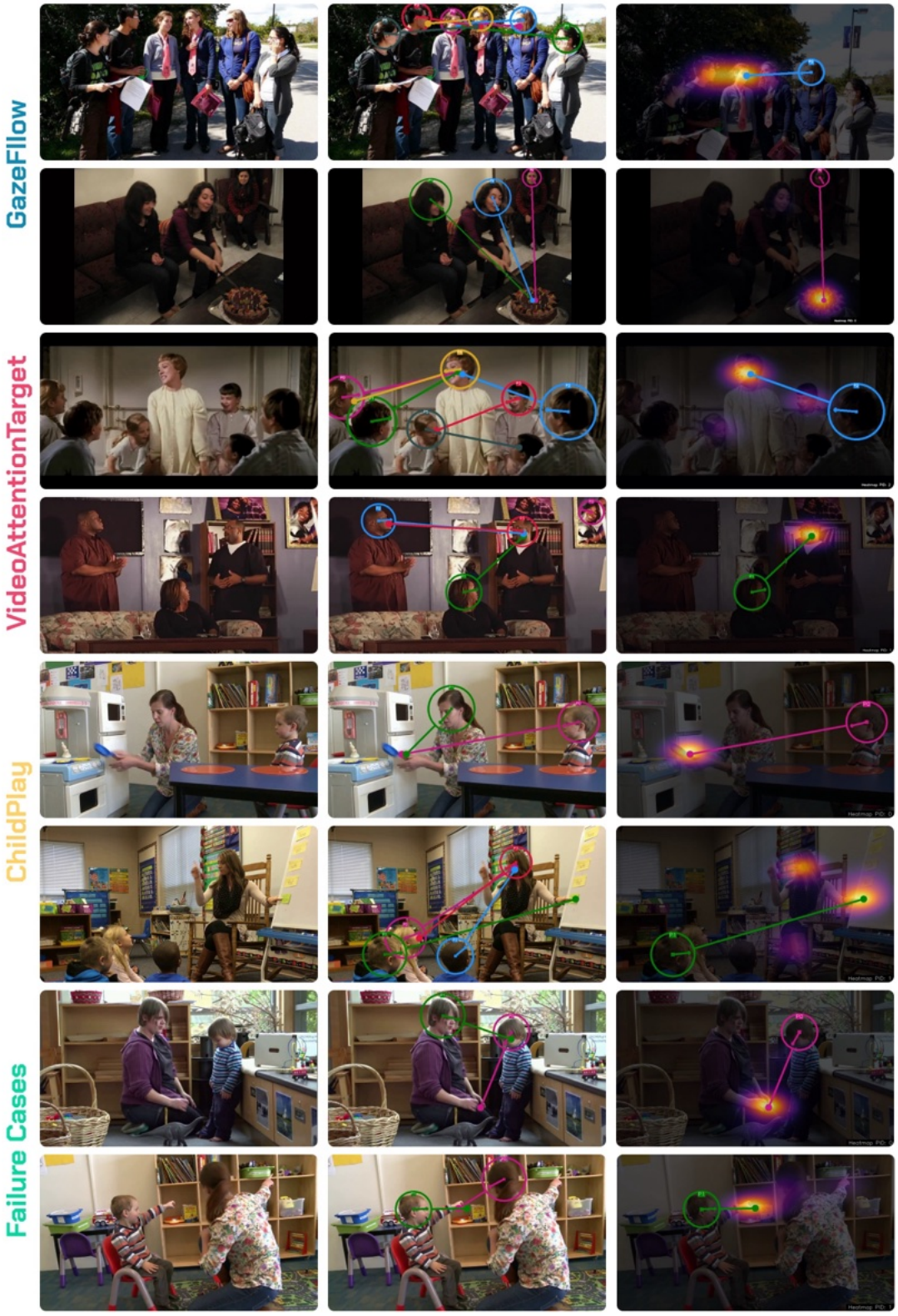


Figure 6.10: Predictions of Sharingan on test images from standard benchmarks. Columns show: input image, point predictions for all people, and the gaze heatmap of a selected person (incorrect predictions shown for failure cases). The model was trained on GazeFollow.

6.7 Conclusion

In this chapter, we introduced Sharingan, a transformer-based architecture designed for efficient and scalable multi-person gaze following. Through extensive experiments, we showed that Sharingan achieves state-of-the-art performance across multiple benchmarks while scaling gracefully with the number of people in the scene. Ablation studies highlighted the importance of person encoding and conditional decoding, and our analysis of output representations shed light on the limitations of coordinate regression compared to heatmap prediction. Despite its strengths, Sharingan still processes scene and gaze tokens in a joint latent space, raising challenges for interpretability, interactions, and scalability in very crowded scenes.

These findings point to the need for going beyond efficiency alone and toward richer modeling of attention in social settings. In particular, gaze in everyday life is rarely an isolated phenomenon: people look at one another, share attention to common objects, and engage in turn-taking behaviors. The next chapter therefore extends Sharingan toward *social gaze behavior*, unifying standard gaze following with higher-level social cues such as joint attention and mutual gaze.

7 Social Gaze Behavior

Social scenes are not just multi-person, they're multi-relational. Beyond locating where each person looks, many applications hinge on interpreting what those "looks" mean: are two people making eye-contact? Is attention jointly coordinated to a common target? Who is looking at whom when turn-taking shifts? In this chapter we extend the multi-person, token-based foundation of the previous chapter to social gaze behavior, introducing UniGaze, a unified model that performs gaze following and predicts key social gaze relations in a single framework. Concretely, we address three behaviors that recur across human interactions: Looking at Heads (LAH; one-way eye-contact), Looking at Each Other (LAEO; mutual gaze or eye-contact), and Shared Attention (SA). We show how to couple person-specific tokens with scene tokens to infer per-person gaze points and pairwise social labels without bespoke pipelines for each task. Figure 7.1 illustrates the difference between traditional gaze target localization approaches and our proposed unified framework, which jointly performs localization and social gaze behavior detection.

Methodologically, this chapter builds on the Sharingan encoder from Chapter 6 by reusing its image-and-person tokenization, while adding custom components to accommodate the social prediction tasks and reason about relations between people. The result is a single architecture that (i) predicts where each person looks and (ii) infers who looks at whom and who shares attention.

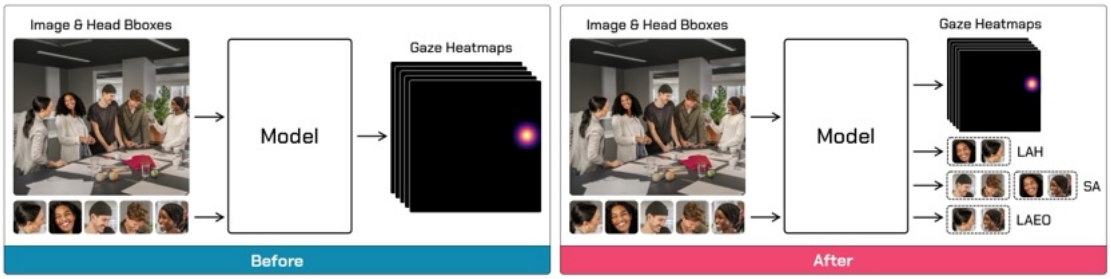


Figure 7.1: Illustration of the difference between previous methods that produce only gaze target locations, and our new method that simultaneously localizes and recognizes the social gaze labels.

At the time of this work (Gupta et al., 2024a), datasets for LAH, LAEO, and SA were fragmented, and we treated each task separately on top of a frozen backbone. A subsequent extension (Gupta et al., 2024b) addressed these limitations by introducing a unified annotation protocol (VSGaze), a novel architecture, and training all tasks end-to-end, enabling stronger cross-task synergy. We briefly revisit these developments as they relate to model architecture, data, and supervision constraints.

7.1 Motivation

From points to social interpretation. Predicting gaze points alone captures *where* someone looks, but not *how those looks connect people*. A predicted heatmap can be numerically close to the ground truth yet fail to reveal whether the gaze lands on another person's head, establishes mutual gaze, or converges with others on a shared target. These distinctions are crucial for interpreting interaction, yet cannot be recovered reliably by post-processing gaze points. Thresholding against head boxes or objects introduces ambiguity when predictions are multi-modal or when gaze lands on "stuff" regions during shared attention. To understand social dynamics, models must learn to interpret gaze jointly with relational labels, rather than treat them as downstream by-products.

Task-specific silos fall short. Past work typically addressed LAEO or SA in isolation, training separate networks on separate datasets. These silos fragment supervision, prevent tasks from reinforcing one another, and struggle to scale beyond dyadic settings. They also impose unrealistic exclusivity: treating "mutual gaze" and "shared attention" as separate, when in practice one person may simultaneously hold eye contact and share attention with another. Without a joint model of gaze and relations, these approaches remain brittle and incomplete.

Why a unified framework? By training gaze following and social relation prediction together, a model can leverage shared cues: eye contact sharpens localization, while accurate low-level gaze predictions disambiguate social states. End-to-end formulations allow such synergy, but until recently supervision was fragmented across datasets with inconsistent labels. The later introduction of unified social annotations and end-to-end training further reduced fragmentation and enabled stronger cross-task synergy, but our focus here is on the first step towards a unified formulation.

Efficiency and scalability still matter. Multi-person efficiency remains central. Social gaze is inherently a group phenomenon, often involving three or more actors in a scene. Our architecture preserves the "encode once, decode many" pattern of Sharingan, reusing the same core representation for both gaze and social predictions. This ensures that scaling from one to many people, and from localization to relations, does not sacrifice throughput.

In summary, this chapter motivates and develops UniGaze—a unified framework for social gaze behavior that: (i) extends gaze following beyond point prediction to include social

relations such as LAH, LAEO, and SA; (ii) overcomes the fragmentation of task-specific silos by leveraging multi-task supervision from heterogeneous data sources and annotations; and (iii) preserves the compute-efficient, multi-person design of the previous chapter, ensuring that gaze predictions are both *socially meaningful* and *practically scalable*.

7.2 Related Work

This chapter focuses on unifying *where* someone looks (gaze following) with *how* those looks relate people (LAH, LAEO, SA). Because a dedicated literature review appears in Chapter 2, we provide a brief, task-focused recap emphasizing aspects most relevant to our formulation and experiments.

7.2.1 Task-Specific Formulation

Early and task-specific lines addressed each social behavior in isolation. Mutual gaze (LAEO) methods framed dyadic eye contact as a binary decision over head pairs (Marin-Jimenez et al., 2014, 2019, 2022), often using geometric cues or pairwise CNNs/Transformers (Doosti et al. (2021); these approaches typically assume temporal evidence, rely on pair enumeration, and do not scale naturally beyond two people. Shared attention (SA) works either combined predicted gaze cones with object proposals or trained binary SA detectors per frame (Fan et al., 2018); such formulations struggled to (i) disambiguate *who shares with whom* when multiple SA groups co-exist, and (ii) identify the specific target instance when labels were frame-level only. The task of LAH emerged later as a semantic probe for gaze following (Tafasca et al., 2023b), but early protocols ignored *which* target head was attended, conflating "looking at any head" with "looking at the correct person". In practice, these silos fragment supervision, prevent cross-task synergy, and require brittle post-processing from gaze points to social labels.

7.2.2 Unified Formulation

A complementary line sought unified treatment. One thread post-processed outputs of strong gaze-following models to infer LAH/LAEO/SA (Tafasca et al., 2024b), benefiting from better localization but remaining sensitive to heatmap multimodality, missed detections, and ambiguous assignments. Another thread proposed joint settings and benchmarks where multiple social behaviors are supervised together (Fan et al., 2019; Chang et al., 2023); these clarified evaluation but often retained limitations such as dyadic enumeration or mutually exclusive "gaze states". Our approach in this chapter takes a middle path: reuse a multi-person gaze backbone once, then *jointly* predict (i) per-person gaze targets and (ii) pairwise social relations in a single forward pass. This avoids bespoke pipelines per task, preserves throughput in crowded scenes, and provides clean pairwise metrics. We note that the later introduction of unified social annotations and end-to-end training further reduced fragmentation and

improved cross-task synergy; those developments are referenced briefly in the discussion but fall outside this chapter's core scope.

7.2.3 Graph Neural Networks

Graph neural networks (GNNs) provide a natural inductive bias for social reasoning: people become nodes, and potential interactions become edges. Message-passing updates let each person aggregate information from relevant others, encouraging globally consistent configurations (e.g. discouraging impossible multi-partner mutual gaze) while remaining permutation-equivariant. Attention-based variants (Veličković et al. (2017), Brody et al. (2021)) are especially appealing here: they learn *which* neighbors matter and by *how much*, enabling the model to parse a scene into a social graph without hand-crafted rules.

Graph-based relational reasoning has also appeared in prior social vision tasks: Fan et al. (2019) use a spatio-temporal GNN on the VACATION dataset to parse mutual gaze, joint attention, and gaze events over time. A follow-up work from de Belen et al. (2023) retains an implicit social graph formulation but adopts a transformer architecture instead. Malik and Isik (2023)'s SocialGNN applies GNNs to infer high-level social interaction categories from relational layouts. Thompson et al. (2021) use message passing to detect conversational groups in multi-person scenes. Meanwhile, Özdel et al. (2024) integrate gaze signals into a scene graph to support action anticipation via GNNs. These works illustrate how graphs can capture person–person and person–scene dependencies. Our design differs in that we fuse gaze localization and social relation prediction in a unified architecture, where we treat the set of person tokens from the gaze backbone as a fully connected graph, and decode relation scores on edges. This strikes a balance between expressivity (explicit relational reasoning) and efficiency (single expensive pass), and integrates seamlessly with multi-task heads for LAH (directed), LAEO (symmetric), and SA (symmetric).

7.3 Architecture

Our architecture builds on the multi-person gaze following model *Sharingan* introduced in the previous chapter, and extends it with new components to predict social gaze behaviors. The system has two main stages: a **Gaze Following Core**, which produces scene-aware person tokens and per-person gaze predictions, and a **Social Gaze Extension**, which leverages those tokens to infer pairwise social relations. Within the extension, a **Relational Graph Block** can be used to explicitly reason about inter-person interactions, and a set of **Multi-Task Social Heads** decode specific social behaviors. Figure 7.2 illustrates the overall design.

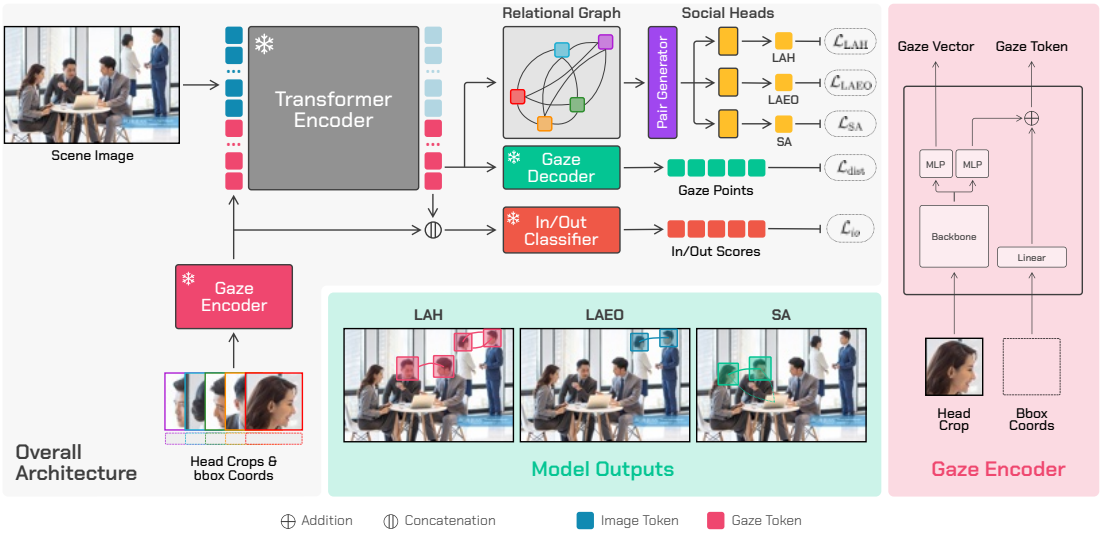


Figure 7.2: Overview of the proposed architecture for gaze following and social gaze prediction. It is composed of two main parts: the *Gaze Following Core* and the *Social Gaze Extension*. Sharingan processes the scene and detected heads to produce image and person tokens, which are jointly encoded by a ViT to yield context-aware person representations. These tokens feed into two branches: (i) a gaze decoder that regresses target locations, and (ii) the Social Gaze Extension, which consists of an optional Graph Attention Network and multiple task-specific heads for predicting social relations (LAH, LAEO, SA).

7.3.1 Gaze Following Core

The backbone of our approach is Sharingan, a hybrid CNN–Transformer architecture designed for multi-person gaze following. It takes as input the scene image, head crops, and their bounding boxes.

Image tokens. The scene image is divided into patches, projected into embeddings, and enriched with positional encodings, following the standard Vision Transformer (ViT) procedure. These embeddings form the *image tokens*.

Person tokens. Each detected head crop is processed by a CNN-based gaze encoder to produce a feature embedding. This embedding serves two purposes: (i) it provides a regressed 2D gaze direction (used by the gaze decoder), and (ii) it is projected into the transformer dimension to form a *person gaze token*. To encode spatial location, the head bounding box is linearly projected and added to the person token.

Joint encoding. Image tokens and person tokens are fed together into a ViT encoder, allowing global attention across people and the scene. The resulting output person tokens encode both individual gaze cues and contextual information. These tokens are used by the gaze following decoders to predict per-person gaze points and in/out labels, and they also serve as input to the Social Gaze Extension.

In this work, we employ an earlier version of Sharingan that regresses gaze coordinates directly

from person tokens using a lightweight MLP, rather than the more advanced *Conditional DPT* decoder introduced later for heatmap prediction. This choice reflects both the maturity of the model at the time and the focus of this chapter: our goal is not to push the frontier of gaze localization itself, but to investigate how gaze following knowledge can be extended to predict social gaze behaviors. As shown in Chapter 6, coordinate regression and heatmap decoding learn similar underlying representations; the difference lies mainly in how the information is extracted at the output. For our purposes, the simpler MLP-based regression suffices to ensure that the model learns a usable internal encoding of gaze localization, allowing us to concentrate on the novel contribution of unified social gaze prediction.

7.3.2 Social Gaze Extension

The extension infers social gaze behaviors—Looking At Heads (LAH), Looking At Each Other (LAEO), and Shared Attention (SA)—directly from the person tokens produced by Sharingan. We consider two variants: (1) passing pairs of tokens directly to task-specific MLP decoders, or (2) first updating tokens with a *Relational Graph Block*.

Relational Graph Block. We implement the relational block as a Graph Attention Network (GAT), which is an attention-based graph neural network architecture introduced in Veličković et al. (2017)). Specifically, we implement our block using GATv2Conv layers from the follow-up work of Brody et al. (2021). Person tokens from Sharingan are treated as nodes in a fully connected graph, where edges represent potential gaze interactions. Each node aggregates information from all others using learned attention weights, updating its representation with socially relevant cues.

Formally, let $\{\mathbf{x}_i\}_{i=1}^{N}$ denote the N person tokens output by Sharingan. The graph is fully connected, so each node i can attend to every other node j. The updated representation $\hat{\mathbf{x}}_i$ of a GATv2Conv layer is given by:

$$\hat{\mathbf{x}}_i = \sigma\Big(\sum_{j=1}^{N} \alpha_{ij}\,\mathbf{W}_t\mathbf{x}_j\Big), \tag{7.1}$$

where the attention weights α_{ij} are computed as

$$e_{ij} = \mathbf{a}^\top \mathrm{LeakyReLU}(\mathbf{W}_s\mathbf{x}_i + \mathbf{W}_t\mathbf{x}_j), \qquad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{N}\exp(e_{ik})}. \tag{7.2}$$

Here $\mathbf{W}_s$, $\mathbf{W}_t$, and $\mathbf{a}$ are learnable parameters, and σ is a non-linearity. This mechanism allows the architecture to model interactions by adaptively weighting which individuals are most relevant when reasoning about each person's social gaze.

Multi-Task Social Heads. After optional graph updates, we form edge representations for every ordered or unordered pair of people by concatenating their token embeddings. These

edge features are processed by lightweight MLP decoders, one per social behavior:

$$e_{i\to j} = \mathscr{D}_{\text{LAH}}\big([\hat{\mathbf{x}}_i \,\|\, \hat{\mathbf{x}}_j]\big), \tag{7.3}$$

$$e_{i\leftrightarrow j} = \mathscr{D}_{\text{LAEO}}\big([\hat{\mathbf{x}}_i \,\|\, \hat{\mathbf{x}}_j]\big), \tag{7.4}$$

$$c_{i,j} = \mathscr{D}_{\text{SA}}\big([\hat{\mathbf{x}}_i \,\|\, \hat{\mathbf{x}}_j]\big), \tag{7.5}$$

where $[\cdot \,\|\, \cdot]$ denotes vector concatenation. In this case, $e_{i\to j}$ is the score that person i is looking at person j (directed relation, LAH), $e_{i\leftrightarrow j}$ is the score that i and j are looking at each other (symmetric relation, LAEO), and $c_{i,j}$ is the score that i and j are sharing attention toward the same target (symmetric relation, SA).

This formulation treats each social behavior as a pairwise classification problem, enabling all tasks to be solved jointly and efficiently in a single forward pass, but requires enumerating all pairs of people in the scene.

7.4 Implementation Details

Loss Function. Training involves a weighted combination of task-specific binary losses:

$$\mathscr{L} = \lambda_{\text{LAH}}\,\mathscr{L}_{\text{LAH}} + \lambda_{\text{LAEO}}\,\mathscr{L}_{\text{LAEO}} + \lambda_{\text{SA}}\,\mathscr{L}_{\text{SA}} + \lambda_{\text{IO}}\,\mathscr{L}_{\text{IO}}, \tag{7.6}$$

where $\mathscr{L}_{\text{LAH}}$ is the loss for Looking at Heads, $\mathscr{L}_{\text{LAEO}}$ for mutual gaze (LAEO), $\mathscr{L}_{\text{SA}}$ for shared attention, and $\mathscr{L}_{\text{IO}}$ for the in/out prediction that determines whether a person's gaze target lies inside the image.

Each of these terms is formulated as a standard binary cross-entropy loss. For the pairwise tasks (LAH, LAEO, SA), the loss is computed for all possible person pairs in the scene and then averaged across pairs. This ensures consistent training even in multi-person settings with variable group sizes.

In principle, an additional regression loss on the gaze target coordinates could be included. However, since the social gaze datasets are relatively small compared to GazeFollow (the dataset used to pretrain Sharingan), we found that adding a target localization loss did not yield improvements in either gaze prediction or social gaze performance. For this reason, we restrict training to the classification losses defined above.

Model Configuration. The Sharingan encoder is pre-trained on GazeFollow Recasens et al. (2015) following the procedure of Tafasca et al. (2024b), and its weights are kept frozen in all experiments. The Relational Graph Block is implemented as a Graph Attention Network (GAT) with two message-passing layers, while the task-specific social decoders are three-layer MLPs with residual connections. The model operates on the full scene image together with cropped

head regions, each resized to 224×224.

Training Procedure. During training, we randomly sample up to six people per scene to enable batch training across varied group sizes. At test time, the batch size is set to one and all detected people in the image are considered. We use the AdamW optimizer Loshchilov and Hutter (2017) with standard weight decay. The learning rate and number of epochs are adapted to each dataset depending on its size, domain, and task complexity. Loss weights are tuned accordingly to balance the contributions of different objectives (LAH, LAEO, SA, and in/out classification). This flexible configuration ensures that optimization remains stable across heterogeneous datasets while preserving comparability in evaluation.

Inference Pipeline. At inference time, the model produces scores for all directed or undirected pairs of people, depending on the task. Since these are formulated as pairwise classifications, we must resolve them into consistent global predictions over the entire scene. We describe our protocol below.

Looking at Heads (LAH). For each person $\mathbf{p}_i$, we obtain scores $\mathbf{e}_{i\to j}$ for all potential targets $\mathbf{p}_j$. To decide the predicted head being looked at, we select the most likely target:

$$\hat{j} = \arg\max_j \ \mathbf{e}_{i\to j}. \tag{7.7}$$

A ground-truth positive LAH instance between $\mathbf{p}_i$ and $\mathbf{p}_{j^*}$ is considered correctly predicted if $\hat{j} = j^*$ and the score $\mathbf{e}_{i\to\hat{j}}$ exceeds the decision threshold (used to build ROC and PR curves). Negative cases are handled analogously: a sample is correct if the top score falls below the threshold.

Looking at Each Other (LAEO). Here a sample is a directed pair of people. Given a person $\mathbf{p}_i$, we compute pairwise scores $\mathbf{e}_{i\leftrightarrow j}$ with all other people $\mathbf{p}_j$ and keep only the highest-scoring partner $\mathbf{p}_{\hat{j}}$. The remaining pairs are then set to 0:

$$\hat{j} = \arg\max_j \ \mathbf{e}_{i\leftrightarrow j}, \qquad \mathbf{e}_{i\leftrightarrow j} = 0 \ \ \forall\, j \neq \hat{j}. \tag{7.8}$$

This avoids inconsistent predictions where a person is simultaneously marked as being in mutual gaze with multiple others.

Shared Attention (SA). For SA, the model outputs probabilities $\mathbf{e}_{i,j}$ for every pair (i, j). Because shared attention is a group phenomenon, we enforce transitivity: if $\mathbf{p}_i$ and $\mathbf{p}_j$ are predicted to share attention, and $\mathbf{p}_j$ and $\mathbf{p}_k$ are also positive, then (i, k) must be considered positive as well. Operationally, we build an undirected graph where edges connect pairs above threshold, then extract the connected components. Each component is interpreted as a group of people sharing attention with one another.

7.5 Experimental Design

7.5.1 Evaluation Protocol

Datasets. We evaluate our approach on four datasets covering complementary aspects of social gaze. These benchmarks were introduced in Chapter 3 and are only briefly summarized here. The VACATION dataset (Fan et al., 2019) provides joint supervision across multiple gaze behaviors, while VideoCoAtt (Fan et al., 2018) focuses on shared attention, and UCO-LAEO (Marin-Jimenez et al., 2019) on mutual gaze.

Since no dataset explicitly annotates the Looking at Heads (LAH) task, we derive labels from the VideoAttentionTarget dataset (Chong et al., 2020b). We first detect all heads in a scene using a strong head detector, and manually verify the results. For each annotated person with a ground-truth gaze point, we then check whether that point falls inside one of the detected head boxes: if so, the instance is labeled positive (the person is looking at that head); otherwise, it is negative. This process yields LAH annotations for VAT, with about 69% of instances labeled positive as shown before in Table 3.2.

We deliberately avoid applying the same derivation to VideoCoAtt or UCO-LAEO, as those datasets only annotate participants engaged in social gaze behaviors. Running the same procedure there would bias the LAH statistics because non-engaged individuals are left unannotated, so negatives would be systematically under-represented, inflating the apparent frequency of positive LAH cases.

Finally, we leverage automatically detected extra heads in all datasets to generate negative pairs wherever possible. For example, if persons A and B are annotated as mutually gazing in UCO-LAEO, and a third person C is detected in the same frame but not labeled, then pairs (A,C) and (B,C) are treated as negative LAEO instances. This strategy ensures more balanced supervision across positive and negative cases without distorting the intended task definitions.

Metrics. For all tasks, we evaluate both *discrimination* and *ranking* performance by computing the ROC curve with its associated AUC score, and the Precision–Recall curve with its associated AP score. These metrics capture how well the model separates positive and negative cases across thresholds.

Shared Attention (SA). In addition to classification quality, we also assess the accuracy of localizing the shared attention target. We report two complementary metrics:

- **Distance:** The ℓ_2 distance between (i) the average of the predicted gaze points of all people involved in a ground-truth SA instance, and (ii) the center of the annotated target bounding box. Distances are normalized to the image scale (1×1).
- **Accuracy:** The proportion of SA instances where the average predicted point falls inside the annotated bounding box. To account for borderline cases, the box is expanded by 5% on each side.

Unlike prior works (Fan et al., 2018; Chong et al., 2020b; Tu et al., 2022), which report SA metrics on a *per-frame* basis, we evaluate on a *per-instance* basis. That is, all participants in a ground-truth SA event are grouped together, and predictions are scored once per SA group rather than once per video frame.

7.5.2 Reference Baselines

We benchmark our approach against existing social gaze methods (Marin-Jimenez et al., 2019; Doosti et al., 2021) and available gaze following models adapted to social gaze prediction through post-processing. Specifically, we evaluate three representative gaze following models: Chong et al. (2020b), Tonini et al. (2022), and Sharingan (Tafasca et al., 2024b). To adapt them to social gaze tasks, we post-process their predicted gaze points:

- **LAH:** A case is positive if the predicted gaze point of $\mathbf{p}_i$ falls inside the ground-truth head box of $\mathbf{p}_j$. Negatives are defined analogously.
- **LAEO:** A pair (i, j) is predicted positive if $\mathbf{p}_i$'s gaze point falls in $\mathbf{p}_j$'s head box *and* vice versa.
- **SA:** A pair (i, j) is positive if the distance between their predicted gaze points is below a threshold. By sweeping thresholds, we compute ROC and PR curves along with AUC and AP scores.

Because post-processing baselines produce binary outputs for LAH and LAEO, they yield only a single precision–recall point rather than a full curve. For a fair comparison, we threshold our models at the highest recall achieved by the post-processing methods and report the corresponding precision values.

Our default method, which we simply denote **UniGaze**, feeds person tokens from Sharingan directly into the task-specific decoders. To test the role of explicit graph-based relational reasoning, we also report the variant **UniGaze (Graph)**, where person tokens are first updated by the Relational Graph Block before being passed to the decoders. Both options are matched in parameter count to ensure fairness.

7.5.3 Benchmark Comparison

VideoAttentionTarget. Results for LAH on VAT are shown in Table 7.4. *UniGaze (Graph)* provides modest gains over *UniGaze* (+1 AUC, +0.6 AP), and outperforms or matches the post-processing baselines of Tonini et al. (2022) and Sharingan. The model from Chong et al. (2020b) remains slightly ahead, as seen in the precision–recall values. It's also worth noting that the low recall value of Sharingan is due to the coordinate regression objective, as explained in Chapter 6. A heatmap variant should, in principle, achieve higher performance compared to the model from Chong et al. (2020b).

VideoCoAtt. Results for SA on VideoCoAtt (Table 7.3) show that *UniGaze (Graph)* achieves the

Model	AUC_{LAEO} ↑	AP_{LAEO} ↑
Marin-Jimenez et al. (2019)	—	0.795
Doosti et al. (2021)	—	0.651
Chang et al. (2023)	—	0.803
UniGaze	**0.986**	**0.957**
UniGaze (Graph)	0.981	0.946

Model	$Prec_{LAH}$ ↑	Rec_{LAH} ↑
Chong et al. (2020b)	0.791	**0.832**
Tonini et al. (2022)	0.790	0.771
Tafasca et al. (2024b)	0.842	0.794
UniGaze	**0.953**	**0.832**
UniGaze (Graph)	0.922	**0.832**

Table 7.1: Results for LAEO on UCO-LAEO.

Model	AP_{LAH} ↑	AP_{SA} ↑	AP_{LAEO} ↑
UniGaze (Graph):LAH	0.924	—	—
UniGaze (Graph):SA	—	0.213	—
UniGaze (Graph):LAEO	—	—	0.721
UniGaze (Graph):ALL	**0.935**	**0.222**	**0.741**

Table 7.2: Results for all tasks (separately and jointly) on VACATION.

Model	AUC_{SA} ↑	AP_{SA} ↑
Chong et al. (2020b)	0.695	0.297
Tonini et al. (2022)	0.715	0.300
Tafasca et al. (2024b)	0.760	0.301
UniGaze	**0.896**	0.594
UniGaze (Graph)	0.890	**0.604**

Table 7.3: Results for SA on VideoCoAtt.

Model	AUC_{LAH} ↑	AP_{LAH} ↑
UniGaze	0.724	0.896
UniGaze (Graph)	**0.733**	**0.902**

Model	$Prec_{LAH}$ ↑	Rec_{LAH} ↑
Chong et al. (2020b)	**0.919**	**0.735**
Tonini et al. (2022)	0.900	0.718
Tafasca et al. (2024b)	**0.919**	0.587
UniGaze	0.875	**0.735**
UniGaze (Graph)	0.890	**0.735**

Table 7.4: Results for LAH on VAT.

Figure 7.3: Comparison of social gaze prediction performance across four benchmark datasets. Best results are indicated in **bold**.

highest AP score and outperforms all post-processing baselines, while *UniGaze* slightly leads in AUC. For target localization metrics, we find that Sharingan yields the best distance (0.116) and accuracy (0.665), outperforming both Chong et al. (2020b) (0.139 / 0.579) and Tonini et al. (2022) (0.127 / 0.452). Because the gaze point decoder is frozen for SA training, localization metrics for UniGaze and UniGaze (Graph) remain identical to Sharingan (Dist.=0.116, Acc.=0.665).

UCO-LAEO. For LAEO on UCO-LAEO (Table 7.1), *UniGaze* achieves the best overall performance, setting a new state-of-the-art, with *UniGaze (Graph)* close behind. As most frames contain only two people, the explicit graph reasoning offers limited additional benefit here. When fixing the recall to match the highest competing value, both of our variants outperform the post-processing baselines on precision by a significant margin.

VACATION. Table 7.2 reports results on the VACATION dataset, where we train *UniGaze (Graph)* both jointly across tasks and separately per task. Please note that this is the first model trained end-to-end on all three tasks. We find that joint training consistently improves performance across LAH, LAEO, and SA, suggesting that multi-task optimization helps the model learn more coherent representations of social interactions.

Gaze Following vs. Social Gaze Metrics. A key observation is that stronger performance on traditional gaze-following metrics does not necessarily translate into better social gaze prediction. For example, although Tonini et al. (2022) achieves superior gaze following performance

compared to Chong et al. (2020b), it performs worse on LAH and LAEO. This highlights that spatial accuracy alone is insufficient for modeling social gaze behaviors, which depend on relational consistency between individuals rather than point-wise localization.

7.5.4 Qualitative Evaluation

Figure 7.4 shows qualitative examples from *UniGaze (Graph)*. The model accurately captures diverse social gaze situations, including cases where faces are partially occluded or not directly visible. However, failure cases occur when subtle ocular or facial cues are missed, leading to incorrect predictions about the presence or absence of social engagement.

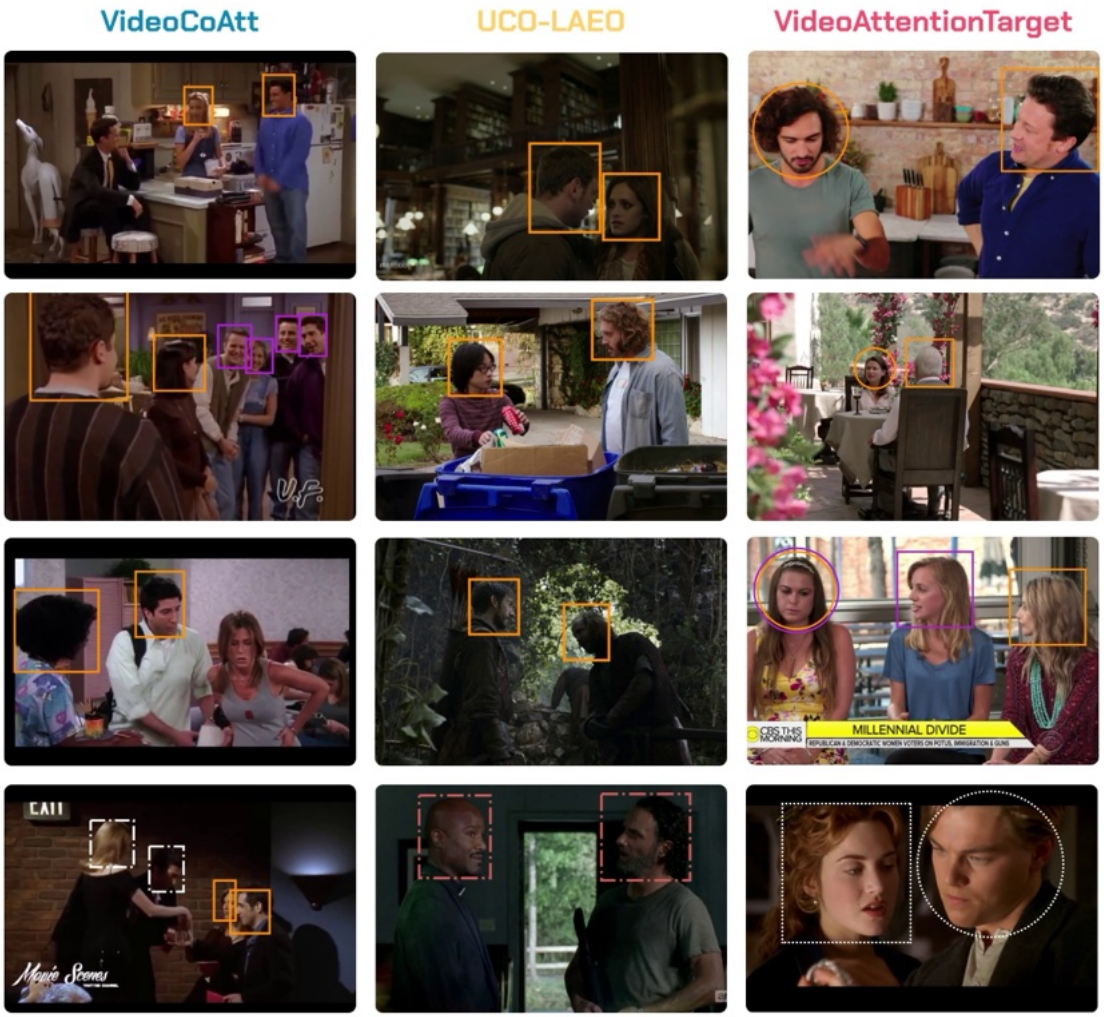


Figure 7.4: Qualitative results on the test sets. The three panels illustrate results for SA (i.e. *VideoCoAtt*), LAEO (i.e. *UCO-LAEO*), and LAH (i.e. *VideoAttentionTarget*), respectively. Shapes of the same color indicate predicted social relationships (e.g., shared attention). True positives are shown in orange and pink (different instances), false negatives in dashed white, and false positives in dashed red. For LAH, the source person is outlined with a bounding box, and the attended head is marked with a circle.

7.6 Limitations

Although *UniGaze* establishes a unified formulation for social gaze prediction, several limitations remain. These limitations arise from constraints in data, modeling assumptions, and inference design, and point to promising directions for future research.

Data scale and annotation completeness. A central bottleneck in social gaze research is the lack of large-scale datasets annotated consistently for all three core tasks—Looking at Heads (LAH), Looking at Each Other (LAEO), and Shared Attention (SA). Existing datasets are small, domain-specific, and often cover only a subset of behaviors, forcing researchers to combine heterogeneous sources or derive missing labels heuristically. This fragmentation prevents joint training and limits generalization to unconstrained scenes. While video datasets provide temporal diversity, they offer little frame-level variability compared to the breadth of image-based collections. The field would benefit greatly from a large, static-image dataset with unified, multi-task social gaze annotations, enabling true joint learning and robust generalization from image diversity rather than redundant video frames. The present work can be viewed as a workaround to this structural bottleneck: an attempt to unify fragmented supervision in the absence of such a dataset.

A partial step toward resolving this was achieved in our follow-up work (Gupta et al., 2024b), which introduced the **VSGaze** dataset. VSGaze unifies gaze points and social gaze annotations across several existing video-based datasets by re-aligning their label spaces and standardizing definitions for Looking at Heads (LAH), Looking at Each Other (LAEO), and Shared Attention (SA). Specifically, it merges annotations from VideoCoAtt, UCO-LAEO, VideoAttentionTarget, and ChildPlay, harmonizing them within a single schema and temporal format while maintaining cross-task consistency. This unified protocol enables end-to-end multi-task training on a single dataset, removing the need for task-specific models. However, while VSGaze represents an important step toward consistent supervision, it still inherits the limited visual diversity of its source video datasets. Furthermore, the missing annotations in each dataset are derived heuristically with limited manual verification. Hence, a large-scale, image-based dataset with unified annotations remains an open need to achieve both consistency and diversity in social gaze learning.

Graph reasoning and implicit message passing. The comparison between *UniGaze* and *UniGaze (Graph)* remains inconclusive. Adding a graph module to update person tokens did not consistently improve performance over directly decoding them, even when parameter counts were matched. This limited gain is consistent with prior findings that Transformer self-attention can itself be interpreted as a dense, fully connected form of message passing (Yun et al., 2019b; Dwivedi and Bresson, 2020; Joshi, 2025). Hence, relational reasoning between people may already be modeled implicitly by the attention mechanism of Sharingan. Nevertheless, explicit graph reasoning might prove beneficial on larger and more diverse datasets where social configurations are more complex.

Architectural decoupling: scene vs. people. Another limitation concerns architectural entanglement. In our design, Sharingan jointly encodes scene and person tokens in a single transformer, and UniGaze optionally adds a person-centric graph block on top. While this yields good performance and preserves the "encode once, decode many" efficiency, it conflates (i) *scene-level* context modeling (layout, saliency, objects) with (ii) *people-level* interaction modeling (who looks at whom), making it harder to (a) inject targeted inductive biases, and (b) control compute allocation across scene vs. people pathways. In practice, the shared latent space can lead to capacity competition between dense scene processing and fine-grained social cues, and the graph head inserted at the end has limited leverage to reshape earlier representations.

A cleaner alternative is a *decoupled, parallel, dual-branch* architecture: a **scene branch** (ViT) processes global image context, a **people branch** models per-person tokens and their relations, and the two exchange information at multiple depths via a *scene–people interaction* module (e.g. cross-attention). This mirrors the philosophy of ViT-Adapter (Chen et al., 2022) for dense prediction—keeping a strong vision backbone while inserting task-specific adapters to route information between streams—and offers several benefits: (1) explicit modularity (we can evolve scene and people reasoning independently), (2) controllable compute (tune capacity per branch and per interaction hop), and (3) clearer places to inject relational priors.

The MTGS follow-up (Gupta et al., 2024b) instantiates this idea: it introduces a custom two-branch design with iterative scene–people exchange and trains the whole system end-to-end with unified social annotations (VSGaze). Empirically, this yields stronger cross-task synergy between gaze following and social labels, suggesting that decoupling the pathways, while allowing them to communicate throughout the network, is a more principled foundation for social gaze than a single mixed latent space with a late graph head.

Pair enumeration and scalability. The proposed method still relies on exhaustive pairwise enumeration to predict social relations. Although lightweight decoders reduce computational overhead, enumerating all possible pairs remains inefficient and unnecessary in scenes with many people, most of whom are not socially engaged with one another. A more scalable approach would involve detecting candidate interaction groups first and applying pairwise reasoning only within these groups. Alternatively, one could adopt an *object-centric formulation* of social gaze: if gaze targets were defined by object or person bounding boxes, social gaze prediction could reduce to computing spatial overlap between gaze and target regions, eliminating the need for exhaustive pair combinations. Such designs would improve both efficiency and robustness to noise in crowded scenes.

Task interdependence. The current framework treats LAH and LAEO as distinct classification tasks, yet LAH is inherently a component of LAEO: mutual gaze arises when two LAH relations are reciprocal. Modeling these dependencies explicitly—either by hierarchical objectives or by joint probabilistic decoding—could yield more coherent and data-efficient learning.

Head representation and spatial precision. The present formulation defines social relations based on whether predicted gaze points fall inside detected head bounding boxes. However, head boxes are imperfect proxies: their size and tightness are subjective, and even well-aligned boxes include background pixels. A more precise formulation would rely on pixel-accurate head masks or segmentation maps, allowing geometric reasoning at the level of visible facial regions rather than coarse boxes. This would make point-in-region evaluations more reliable and less sensitive to annotation style.

Inference design. Our current inference protocol relies on hard argmax assignments or thresholded pairwise decisions to infer global scene relations. Such discrete rules can produce inconsistent or unstable predictions, especially in multi-person settings where probabilities are close. A more principled alternative is to keep soft pairwise scores and perform a global, scene-level optimization to identify interaction groups—e.g., using maximum-weight matching for LAEO or correlation clustering for SA. This approach would yield predictions that are consistent with relational constraints (symmetry, transitivity) and less sensitive to arbitrary thresholds.

In summary, the limitations of *UniGaze* highlight broader challenges for the field: the scarcity of unified data, the need for scalable and structure-aware formulations, and the importance of robust, globally consistent inference. Addressing these aspects will be crucial for moving from pairwise social gaze detection toward comprehensive models of group attention and social dynamics.

7.7 Conclusion

In this chapter, we introduced *UniGaze*, a unified framework for social gaze prediction that integrates standard gaze following with higher-order relational cues such as Looking at Heads (LAH), Looking at Each Other (LAEO), and Shared Attention (SA). Building on the multi-person transformer backbone of Sharingan, UniGaze extends gaze understanding from individual localization to social interaction reasoning. Through systematic comparisons across multiple benchmarks, we demonstrated that UniGaze achieves competitive or state-of-the-art results while maintaining scalability and efficiency. Joint multi-task training further improved performance across tasks on the VACATION dataset, suggesting that shared representations of gaze and social behavior reinforce one another.

At the same time, our analysis revealed several important insights. First, explicit graph-based reasoning provided limited gains over the transformer baseline, supporting the view that self-attention already performs dense message passing among person tokens. Second, the study exposed practical challenges in data availability and task formulation: existing datasets remain small, fragmented, and inconsistently annotated, limiting both generalization and the potential of multi-task learning. Finally, we highlighted broader modeling and inference issues, including the inefficiency of exhaustive pair enumeration, the subjectivity of bounding-box

definitions, and the need for globally consistent scene-level reasoning.

Together, these findings show that social gaze understanding is a crucial intermediate step between low-level gaze localization and high-level semantic interpretation. It bridges the gap between *where* people look and *how* these looks connect them, paving the way for models that not only locate attention but also understand its meaning. The next chapter therefore extends this direction toward *semantic grounding*, moving from social relations to the identification of gaze targets in the scene: linking visual attention to object semantics and open-vocabulary understanding.

8 Semantic Grounding

In this chapter, we move from social semantics to object semantics, asking not only *where* attention lies or *how* it is shared among people, but also *which* entities in the environment are being attended to. In everyday scenes like games, clinics, kitchens, and streets, knowing *what* someone is attending to matters as much as *where* they are looking. A single pixel or heatmap peak is informative but incomplete for downstream reasoning: a robot needs to grasp the cup, a clinician cares that a child attends to a person rather than the screen, and a retail analyst distinguishes shelf from product. Building on the previous chapter's focus on social gaze behavior, which added meaning through interpersonal relations, we now explore semantics from a complementary perspective: grounding gaze in the objects and regions of the scene. This chapter develops semantic gaze target prediction: jointly localizing the attended region and assigning it a class label, so that gaze becomes a bridge from pixels to meaning. We frame recognition as vision–language alignment to support large, flexible vocabularies rather than fixed detectors, keeping the door open to new categories at inference. Figure 8.1 illustrates the contrast between traditional gaze target localization approaches and our proposed framework, which jointly performs localization and recognition.

8.1 Motivation

From points to meaning. Classic gaze following predicts a point/heatmap from a scene branch

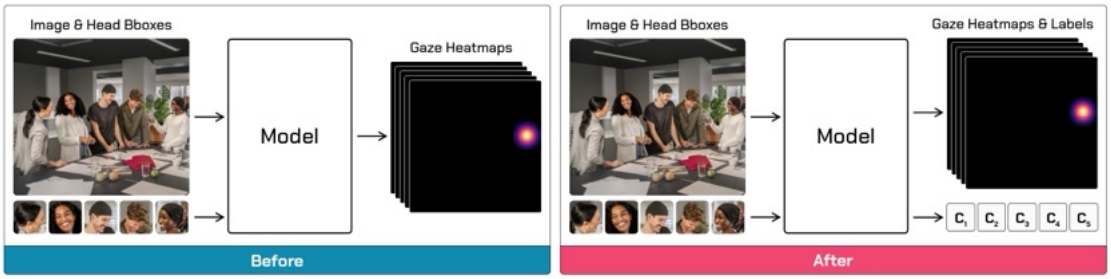


Figure 8.1: Illustration of the difference between previous methods that produce only gaze target locations, and our new method that simultaneously localizes and recognizes the gaze target.

and a head branch, which is powerful, general, and equipment-free, but insufficient whenever decisions depend on the identity of the target (object/person/stuff). Post-processing a gaze point with an object detector (i.e. checking if a gaze point falls inside an object's bounding box) adds compute and fragility, struggles with unbounded "stuff" categories, and raises ambiguous matching questions when heatmaps are multi-modal or points fall into multiple boxes. Jointly training localization *and* target class lets the model learn how to interpret its own heatmaps to recover the right category, avoiding brittle cascades.

Gaze following as human-object interaction. Framing "looking" as a verb in a human-object interaction task seems tempting, but existing datasets rarely label looking systematically; where they do (e.g. V-COCO), coverage is limited to 80 classes and biases toward a handful of stereotyped targets (e.g. TV, phone). This encourages shortcut learning—detecting a TV almost implies "watching"—and under-represents the long tail of everyday gaze targets, including "stuff." We therefore introduce dedicated benchmarks/protocols that couple gaze localization with rich, open vocabulary labels.

Why open-vocabulary recognition? Real deployments face shifting label spaces across domains and users. We cast recognition as aligning a learned gaze label embedding to text embeddings from a frozen language encoder, allowing arbitrary vocabularies at inference and reducing dependence on closed, task-specific detectors. This also supports hierarchical or synonymous labels more naturally than fixed softmax heads.

Multi-person efficiency remains central. Multi-person scenes are the *norm*, not the *exception*. Our formulation preserves the "encode once, decode many" pattern from the previous chapters, so adding semantics doesn't sacrifice throughput, which is crucial for a wide range of applications (e.g. real-time robotics, large-scale clinical screening). In short: *useful* gaze is both efficient and semantic.

In summary, this chapter motivates and develops a semantic gaze target prediction framework that: (i) extends gaze following beyond point localization to include open-vocabulary recognition of attended targets; (ii) overcomes the limitations of post-hoc detector matching, especially for "stuff" categories and ambiguous large objects; and (iii) preserves the compute-efficient, multi-person design introduced earlier, thereby making gaze predictions both meaningful and practical for downstream applications.

8.2 Related Work

We situate semantic gaze target prediction at the intersection of gaze following, human–object interaction (HOI), and vision–language modeling (VLM). Classical gaze-following methods localize "where someone is looking" as a point or heatmap, without naming the target. HOI detects ⟨human, verb, object⟩ triplets. VLMs align visual tokens and natural language. Our contribution builds a bridge: from non-semantic gaze localization to open-vocabulary, text-

grounded gaze targets, using ingredients and evaluation practices drawn from all three areas.

8.2.1 Semantic Gaze Following

To the best of our knowledge, no prior work has directly tackled the problem of semantically grounding gaze target prediction. The closest attempt is the object-aware framework of Tonini et al. (2023), which couples object detection with gaze prediction and employs a gating mechanism to assign a bounding box and class label to the gaze target *if* a suitable object is detected along the line of sight. However, this formulation inherits the limitations of standard detectors: when an object is missed or falls outside the detector's vocabulary, no semantic output is produced. Functionally, this reduces to post-processing a predicted gaze point against detector boxes, since no gradient flows from the recognition task back into the gaze representation. The main benefit over naive cascades is the shared backbone, which amortizes feature extraction for both gaze localization and general-purpose object detection.

In parallel, the GOO benchmark (Tomas et al., 2021) introduced "gaze object prediction" with bounding-box supervision in retail scenes, but as explained in Chapter 3, its setting is narrow (retail shelves) and heavily synthetic, limiting generalization and the scope of semantics.

8.2.2 Human-Object Interaction

HOI detection aims to localize people and objects and classify their interaction (e.g. ⟨person, ride, bicycle⟩), typically on V-COCO (Gupta and Malik, 2015) and HICO-DET (Chao et al., 2018). V-COCO annotates people with a small verb set and role boxes; HICO-DET scales to 600 HOI categories (117 verbs and 80 objects), defining interaction classes like ⟨watch, tv⟩. These benchmarks clarify the task and evaluation but bind "watch/look" to specific object lists rather than an open notion of "gazed target." On the other hand, HOI methods often rely on inductive biases that suit physical interactions such as person-object spatial layout encoding (Gao et al., 2018; Gkioxari et al., 2018). While useful when touch/proximity dominate, they are misaligned with gaze, where targets may be distant and disjoint from hands or contact. In short, treating gaze as a generic HOI loses person-specific directional cues (head/eyes), cone-aware field of view, and scene-level saliency priors that are central to gaze reasoning. Our formulation preserves those gaze-specific inductive biases while adding semantic grounding.

8.2.3 Vision-Language Modeling

VLMs align image and text embeddings with contrastive and matching objectives, enabling zero-shot naming and open-vocabulary recognition. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) established scalable dual-encoder pretraining on web data; subsequent models like ALBEF (Li et al., 2021a), BLIP (Li et al., 2022a), and BLIP-2 (Li et al., 2023) combine contrastive alignment with cross-attention or lightweight bridging to LLMs for stronger grounding and transfer. These tools motivate our use of language supervision to *name* gaze targets without

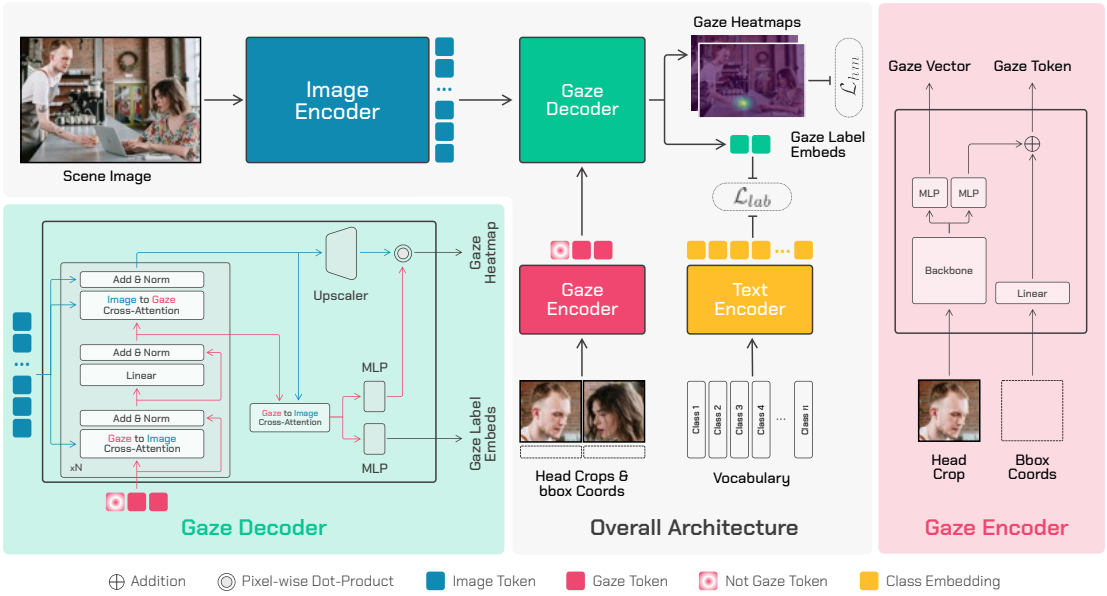


Figure 8.2: Overview of our proposed SemGaze architecture. (i) The scene image is processed by the *Scene Encoder*, producing a set of scene tokens (blue squares). (ii) Each person is represented by a *Gaze Token*, obtained from their head crop and bounding box via the *Gaze Encoder* (orange squares). (iii) Scene and gaze tokens interact within the *Gaze Decoder*, where two-way cross-attention iteratively aligns them and produces both gaze heatmaps and (normalized) gaze label embeddings (yellow squares). (iv) In parallel, a frozen *Text Encoder* maps a vocabulary of class names into (normalized) text embeddings (green squares). (v) The predicted gaze label embeddings are compared with the text embeddings via similarity, yielding the final class prediction for each gaze target.

fixing a closed label set, and to evaluate open-world generalization.

In summary, prior gaze works produce where (points/heatmaps) but are silent about the identity of the target; HOI detects *what–verb–what* but lacks gaze-specific priors and open "looking at X" semantics; VLMs provide the missing alignment. This chapter is, to our knowledge, the first treatment to unify these threads into a principled task definition, architecture, and evaluation protocol for semantic, language-grounded gaze target prediction, positioning this research as a foundation for future benchmarks and methods.

8.3 Architecture

Our design is inspired by the idea of promptable architectures, as in promptable segmentation (Kirillov et al., 2023), where a single image representation can be conditioned by different prompts to produce diverse outputs. We extend this idea to gaze following by separating scene encoding from person-specific prompting. Instead of jointly processing head and image tokens, as in Sharingan (Tafasca et al., 2024b), we first encode the scene once in a person-agnostic manner and then prompt it with lightweight gaze tokens derived from head crops and bounding boxes (see Figure 8.2).

This separation is motivated by two shortcomings of the previous Sharingan architecture.

First, its joint token processing reduces interpretability, as gaze tokens and scene tokens intermingle despite belonging to distinct modalities. Second, scaling remains marginally tied to the number of people, since each additional gaze token is appended at the input of the main transformer encoder, where attention scales quadratically with sequence length. By contrast, our promptable design cleanly and completely decouples the heavy computation from the number of people, enabling an *encode-once, decode-many* paradigm that is both efficient and transparent.

At the same time, we preserve valuable ideas from Sharingan: the controllable gaze token derived from a high-resolution head crop, the positional grounding of gaze tokens via head bounding boxes, and the principle of expensive encoding followed by lightweight decoding. Together, these choices yield an architecture that is elegant, scalable, and faithful to the structure of the gaze problem. Intuitively, the image encoder learns to highlight candidate gaze targets (e.g. salient objects, people, or "stuff"), while the decoder acts as a filtering mechanism that selects the correct target based on the query person. We return to this hypothesis in section 8.5.4.

We now describe the main components of the architecture.

8.3.1 Scene Encoder

Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, we employ a transformer encoder to produce a sequence of image tokens $\mathbf{x}_s \in \mathbb{R}^{N \times D}$, where $N = \frac{W}{P} \cdot \frac{H}{P}$ is the number of patches, P is the patch size, and D is the embedding dimension. This design ensures the encoder is person-agnostic: the entire scene is processed once, independent of how many gaze queries will be decoded later. While we adopt a transformer by default, a convolutional backbone is also possible, provided that the output feature map is equipped with positional encodings.

8.3.2 Gaze Encoder

The gaze encoder is identical to the one introduced in Sharingan (Tafasca et al., 2024b). Its role is to capture person-specific directional cues by processing each head crop at high resolution (e.g. 224×224). Given a head crop $\mathbf{h}_{\text{crop}} \in \mathbb{R}^{h \times w \times C}$, we pass it through a backbone followed by a two-layer MLP to produce: a gaze token $\mathbf{x}_g \in \mathbb{R}^D$, used to prompt the decoder. The gaze encoder also outputs a gaze vector $\mathbf{g}_v \in \mathbb{R}^2$, supervised with an angular loss for auxiliary orientation supervision.

To make the token spatially aware, we project the corresponding head bounding box $\mathbf{h}_{\text{bbox}} \in [0,1]^4$ to $\mathbf{x}_{\text{bbox}} \in \mathbb{R}^D$ and add it to the gaze token. For M heads in an image, we obtain $\mathbf{x}_g \in \mathbb{R}^{M \times D}$.

8.3.3 Gaze Decoder

The goal of the gaze decoder is to integrate the person-specific prompts (gaze tokens) with the scene representation (scene tokens) in order to predict both gaze heatmaps and gaze label embeddings. Conceptually, the decoder aligns the two streams of information: the scene encoder highlights candidate regions in the image, while the decoder filters and grounds them according to the query person.

The decoder is implemented as a stack of B transformer blocks, each composed of two cross-attention layers (gaze-to-image and image-to-gaze) and a feed-forward network, with residual and normalization layers in between. This two-way design ensures that gaze tokens and image tokens exchange information symmetrically.

Two-way Cross-Attention. Let $\mathbf{x}_g \in \mathbb{R}^{M\times D}$ be the M gaze tokens and $\mathbf{x}_s \in \mathbb{R}^{N\times D}$ the N scene tokens. A generic cross-attention operation can be written as:

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{8.1}$$

where Q are the queries, K the keys, and V the values (Vaswani et al., 2017).

We use two complementary directions:

$$\tilde{\mathbf{x}}_g = \mathrm{Attn}\big(\mathbf{x}_g W_Q, \mathbf{x}_s W_K, \mathbf{x}_s W_V\big), \tag{8.2}$$

$$\tilde{\mathbf{x}}_s = \mathrm{Attn}\Big(\mathbf{x}_s W'_Q, \mathbf{x}_g W'_K, \mathbf{x}_g W'_V\Big). \tag{8.3}$$

The first operation ($\mathbf{x}_g \to \mathbf{x}_s$) lets gaze tokens query the scene and gather contextual information, while the second ($\mathbf{x}_s \to \mathbf{x}_g$) updates the scene representation with person-specific information. Stacking both directions across N blocks enables iterative alignment between the two modalities.

Prediction Heads. After the final block, we obtain updated gaze tokens $\mathbf{x}_g \in \mathbb{R}^{M\times D}$ and scene tokens $\mathbf{x}_s \in \mathbb{R}^{N\times D}$. These are fed into two task-specific heads:

(i) **Gaze label head:** a 6-layer MLP maps each gaze token to a label embedding,

$$\mathbf{x}_l = \mathrm{MLP}_{\mathrm{label}}(\mathbf{x}_g) \in \mathbb{R}^{M\times D_\ell}, \tag{8.4}$$

which is aligned with text embeddings for open-vocabulary recognition.

(ii) **Gaze heatmap head:** a 3-layer MLP projects each gaze token to a reduced dimension,

$$\hat{\mathbf{x}}_g = \mathrm{MLP}_{\mathrm{hm}}(\mathbf{x}_g) \in \mathbb{R}^{M\times d}, \quad d < D. \tag{8.5}$$

In parallel, the image tokens are reshaped into a spatial feature map, upsampled via the

Upscaler module to the resolution $(H_{\text{hm}}, W_{\text{hm}})$, and channel-reduced:

$$\hat{\mathbf{x}}_s \in \mathbb{R}^{H_{\text{hm}} \times W_{\text{hm}} \times d}. \tag{8.6}$$

Finally, for each projected gaze token, we apply a dot-product with every spatial position of the upscaled scene tokens to obtain the heatmaps:

$$\mathscr{H}_k(i,j) = \hat{\mathbf{x}}_g^{k\top} \cdot \hat{\mathbf{x}}_s(i,j), \quad k = 1,\ldots,M. \tag{8.7}$$

This dot-product acts as a filtering mechanism, localizing the gaze target by aligning gaze tokens with spatial positions in the scene.

The Background Token. In the image-to-gaze cross-attention, each image token is updated as a weighted sum of the gaze tokens. Without a safeguard, even irrelevant background regions are forced to mix with person tokens. To mitigate this, we introduce a learnable *not-gaze token* that serves as a "background" option. Conceptually, it allows background regions to attend to a neutral representation rather than spuriously combining with real gaze tokens, thereby improving robustness and interpretability.

In/Out Prediction. While this chapter focuses exclusively on gaze target localization and recognition, the proposed architecture can be readily extended to also predict whether the gaze target lies inside or outside the image. This requires only a minor modification: we add an additional MLP head, structurally similar to the label embedding head, that maps the output gaze tokens $\mathbf{x}_g$ to a scalar probability. Formally,

$$p_{\text{in}} = \sigma\big(\text{MLP}_{\text{inout}}(\mathbf{x}_g)\big), \tag{8.8}$$

where σ denotes the sigmoid function. The output $p_{\text{in}} \in [0,1]$ can then be interpreted as the probability that the gaze target is located inside the image. This design preserves the modularity of the architecture: each prediction task (heatmap, label, in/out) is handled by a dedicated lightweight MLP, while sharing the same scene and gaze representations.

8.3.4 Text Encoder

Since we formulate gaze target recognition as a vision–language alignment problem, our architecture requires a text encoder to map class labels into a shared embedding space. For this purpose, we adopt the CLIP text encoder (Radford et al., 2021), which we keep frozen throughout training. This choice leverages the broad coverage and open-vocabulary capabilities of CLIP without introducing additional training complexity.

During training, the predicted gaze label embeddings are aligned with the corresponding text embeddings of the ground-truth classes. At inference time, we first define a vocabulary of candidate class labels, convert them into text embeddings using the frozen encoder, and then compute their similarity with the predicted gaze label embedding. The final predicted

class corresponds to the label with the highest similarity score. This design enables open-vocabulary recognition, allowing our model to generalize to novel categories beyond those seen during training.

8.4 Implementation Details

Loss Function. We optimize the model with a weighted sum of three objectives: (i) a pixel-wise MSE loss for heatmap prediction, (ii) an angular loss for regressing the normalized gaze vector, and (iii) a contrastive loss for the gaze label embedding.

The label loss follows the general structure of the multimodal InfoNCE contrastive objective used in CLIP (Radford et al., 2021), but with two important modifications tailored to our setting. First, we only consider the *image-to-text* direction, since the text encoder is frozen and provides stable supervision. Second, because our labels correspond to (pseudo-)class names rather than unique captions, multiple samples in a batch may share the same class. To avoid penalizing such redundancy, we compute the denominator over the set of *unique* classes in the batch.

Formally, let $\mathbf{x}_i^l \in \mathbb{R}^{D_\ell}$ denote the predicted gaze label embedding for sample i, and $\mathbf{x}_c^t \in \mathbb{R}^{D_\ell}$ the text embedding of class c. Each sample i is associated with a ground-truth class y_i. Let $\mathscr{C}$ be the set of unique classes in the batch:

$$\mathscr{C} = \{ y_i \mid i = 1, \ldots, N \}.$$

The label loss is then defined as

$$\mathscr{L}_{lab} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(s(\mathbf{x}_i^l, \mathbf{x}_{y_i}^t)/\tau\right)}{\sum_{c \in \mathscr{C}} \exp\left(s(\mathbf{x}_i^l, \mathbf{x}_c^t)/\tau\right)}, \tag{8.9}$$

where $s(\cdot,\cdot)$ is the cosine similarity and τ is a learnable temperature.

This objective encourages each predicted gaze label embedding to be close to the corresponding class embedding while being pushed away from all other unique classes in the batch. The formulation is closely related to supervised contrastive learning (Khosla et al., 2020), but differs in that we explicitly collapse duplicates of the same class to a single anchor in the denominator. This avoids introducing spurious negatives when multiple samples share the same gaze label. Moreover, by aligning predictions with a fixed text embedding space, the model learns to generalize to open-vocabulary settings beyond the training classes.

Model Configuration. The scene image is processed at 256×256 and the head crop at 224×224, producing a 64×64 gaze heatmap and a 512-dimensional label embedding. The decoder uses a feature dimension of $d = 96$, two transformer blocks, and is trained with $N_p = 1$ person per image. The text encoder is CLIP (Radford et al., 2021), kept frozen. The gaze encoder is

based on ResNet-18 pretrained on Gaze360 (Kellnhofer et al., 2019), and the scene encoder is ViT-Base initialized from Multimodal MAE (Bachmann et al., 2022), both of which are identical to Sharingan.

Training Procedure. For GazeFollow, we train with AdamW (lr=2×10^{-4}, weight decay 0.003), a cosine annealing schedule with 4 warmup epochs, and a total of 20 epochs. The batch size is 300, and the loss weights are $\lambda_{hm} = 1000$, $\lambda_{lab} = 1$, and $\lambda_{ang} = 3$. We apply stochastic weight averaging (Izmailov et al., 2018) for improved generalization. On GazeHOI, we remove the warmup and reduce the learning rate to 2×10^{-5}. All experiments run on either a single RTX 3090 (24GB) or H100 (80GB), with training times ranging from 2–10 hours depending on dataset size.

When training on GazeHOI, the object box center is used as a proxy for the gaze point, but this is suboptimal for large objects (e.g. the center of a person's box lies near the hips rather than the head). To address this, if the object is larger than a threshold, we instead use the prediction of a GazeFollow-trained model whenever it falls inside the ground-truth box; otherwise, we default to the box center.

Inference Pipeline. At inference, we first construct a vocabulary of class names (either from the dataset or defined externally) and encode them with the frozen text encoder to obtain text embeddings. The text encoder can then be discarded, as label prediction reduces to comparing the predicted gaze label embedding with the vocabulary embeddings and selecting the top-k classes by similarity score. Additionally, instead of using the standard argmax to extract gaze points from heatmaps, we adopt the Distribution-Aware coordinate Representation of Keypoints (DARK) decoding method of Zhang et al. (2020).

8.5 Experimental Design

8.5.1 Reference Baselines

In addition to comparisons with prior gaze target localization methods, we introduce a strong two-stage baseline for gaze target recognition, since no such approaches exist in the literature. The baseline operates as follows. First, we freeze our proposed gaze model and use it to predict a gaze heatmap. This heatmap is then used to condition the input image by emphasizing the attended region, after which CLIP's pretrained vision encoder (Radford et al., 2021) produces the corresponding gaze label embedding. In this way, the baseline leverages the alignment between CLIP's vision and text encoders for open-vocabulary recognition.

We consider three conditioning strategies for incorporating the predicted gaze heatmap into the image: (i) masking, (ii) blurring, and (iii) cropping. Figure 8.3 illustrates the baseline design and the three variants. An additional advantage of this setup is that the baseline's recognition accuracy depends on the same predicted gaze locations as our model, allowing us to control for localization quality and perform a fair, unbiased comparison of recognition performance.

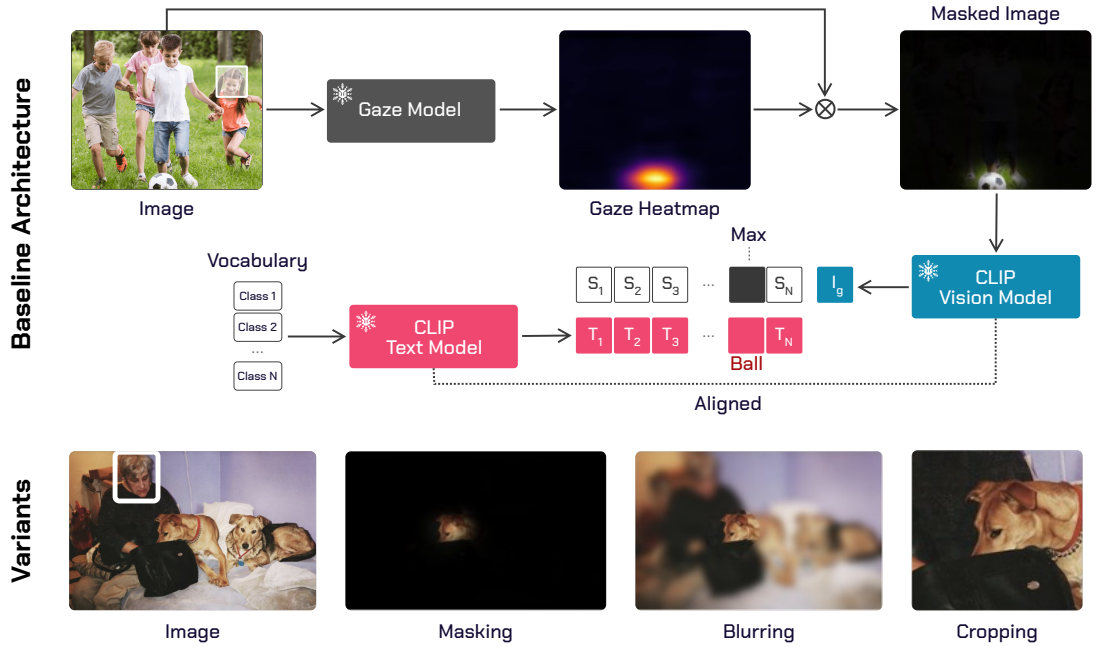


Figure 8.3: Top: Overview of the proposed baseline architecture with masked conditioning. Similarity scores S_i are computed between the predicted gaze label embedding $\mathbf{I}_g$ and the class embeddings $\mathbf{T}_i$. **Bottom:** Visualization of the different conditioning strategies applied using the predicted gaze heatmap.

Since CLIP has likely not been trained on images that are heavily masked, blurred, or cropped, we introduce two additional baseline variants that adapt its vision encoder more directly. (1) We fine-tune the crop-based baseline on GazeFollow, using crops derived from the ground-truth heatmaps during training rather than predicted ones. (2) Instead of modifying the image input, we apply the conditioning at the feature level: the image is passed unaltered through CLIP's vision encoder, and we then compute a weighted average of the resulting image tokens using the (downscaled) predicted heatmap as weights, followed by a projection layer. This second variant allows the model to retain surrounding context even when the heatmap highlights only part of the attended object.

8.5.2 Benchmark Comparison

We report quantitative results on the GazeFollow dataset in Table 8.1. For localization, our architecture establishes a new state-of-the-art across both metrics, surpassing the second-best method (i.e. Sharingan) by 4.4% and the third-best of Jin et al. (2021) by 8.5% on Avg. Dist. For recognition, our method significantly outperforms all three frozen baseline variants, improving Acc@1 by more than 20 percentage points compared to the next best (the crop variant). By contrast, the two fine-tuned baselines close the gap: the crop-based version performs slightly below our model, while the heatmap-weighted version performs slightly above.

Two points are important when interpreting these baselines. First, they build directly on

Method	Localization		Recognition		
	Avg. Dist. ↓	Min. Dist. ↓	Acc@1 ↑	Acc@3 ↑	SetAcc@1 ↑
Random	0.471	0.391	0.002	0.010	0.003
Bias / Majority	0.295	0.229	0.010	0.015	0.011
Chong et al. (2020b)	0.137	0.077	—	—	—
Jin et al. (2021)	0.126	0.076	—	—	—
Fang et al. (2021)	0.124	0.067	—	—	—
Jin et al. (2022)	0.118	0.063	—	—	—
Bao et al. (2022)	0.122	—	—	—	—
GeomGaze	0.125	0.064	—	—	—
Sharingan	<u>0.113</u>	<u>0.057</u>	—	—	—
Baseline (Mask)	0.108	0.051	0.124	0.253	0.147
Baseline (Blur)	0.108	0.051	0.190	0.362	0.222
Baseline (Crop)	0.108	0.051	0.239	0.428	0.278
Baseline (Crop FT)	0.108	0.051	0.437	0.588	0.504
Baseline (Hm weight FT)	0.108	0.051	**0.466**	**0.653**	**0.542**
Baseline† (Hm weight FT)	0.137	0.077	0.442	0.620	0.514
SemGaze	**0.108**	**0.051**	<u>0.447</u>	<u>0.642</u>	<u>0.516</u>

Table 8.1: Results on the GazeFollow dataset. The best scores are shown in bold and the second-best are underlined. The term *FT* means that the model is fine-tuned, while *Hm weight* refers to the heatmap weighted baseline. All baseline variants rely on our model for gaze heatmap prediction, except for Baseline†, which uses the model from Chong et al. (2020b).

our model's localization, which is already the most accurate to date. To confirm this, we replaced the gaze heatmap predictor in the best-performing baseline with the model from Chong et al. (2020b), and performance dropped below that of our method (*cf.* the row of Baseline†). Second, the baselines include an additional vision transformer pretrained on large-scale recognition data, whereas our model uses only a lightweight MLP head (3.3M parameters) for label prediction. Overall, our method has 116M parameters compared to 200M for the baseline (i.e. a 42% reduction), while remaining competitive in recognition and clearly superior in localization. We further show in ablations (*cf.* Section 8.5.3) that forcing CLIP's vision encoder to perform both localization and recognition degrades performance, highlighting why the baseline requires two specialized ViTs.

Finally, manual inspection of test samples where our model and the best heatmap-weighted baseline disagree reveals that most errors are due to hierarchy or semantic similarity (e.g. predicting *ball* instead of the ground-truth *racket*). Refer to Figure 8.4 for some illustrative examples.

On GazeHOI (Table 8.2), we observe a similar trend. Our method outperforms the zero-shot baselines by a wide margin, while both fine-tuned variants achieve higher recognition scores. We note that a GazeFollow pretrained model already performs well on localization in GazeHOI (65% vs. 72% Gaze Accuracy), but its recognition performance is much weaker than the fine-tuned version (30% vs. 58% Acc@1). This gap is likely due to vocabulary mismatch: around 150 GazeHOI classes are absent from GazeFollow pseudo-labels.

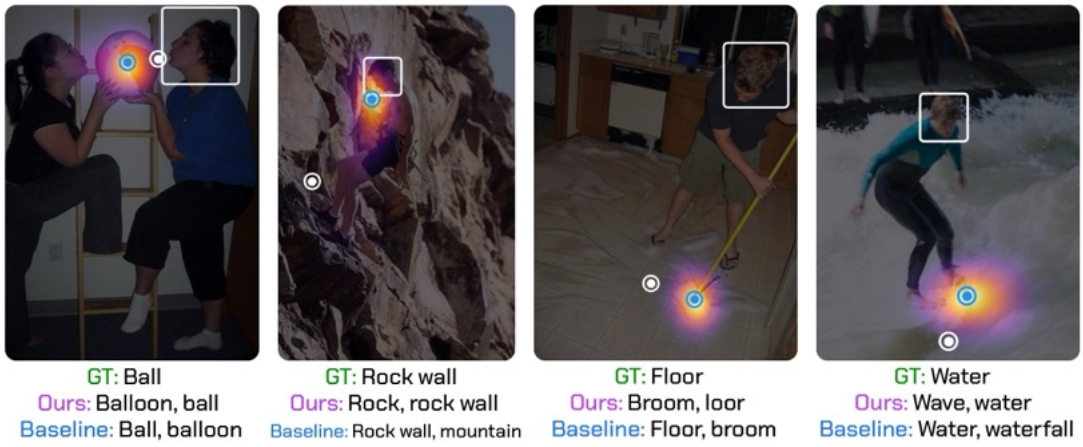


Figure 8.4: Comparison with the baseline on ambiguous cases. In these examples, the baseline prediction matches the ground truth, while our model is marked incorrect despite being semantically plausible. The annotated person is shown in white, the predicted gaze point in blue, and the predicted heatmap overlaid on the image.

It is also important to emphasize that, unlike the baselines, our gaze label prediction head is trained entirely from scratch. Given the relatively small scale of current datasets (100K instances vs. 400M in CLIP pretraining), we believe dataset size is a limiting factor in recognition performance.

8.5.3 Ablation Studies

Batch Size. Contrastive learning methods are known to benefit from larger batch sizes (Radford et al., 2021; Chen et al., 2020), as more negatives provide stronger training signals. We vary the batch size from 48 to 300 and observe steady improvements up to the high end, where performance begins to plateau (*cf.* Table 8.3). This suggests that beyond a certain point, additional negatives yield diminishing returns under our current setting and dataset size. Localization performance on the other hand, remains stable.

Losses. We ablate the contribution of the different training losses (*cf.* Table 8.4). Training with only the label loss, without explicit localization supervision, reduces Acc@1 by 3.4%, showing that recognizing *what* a person is looking at is significantly easier once the model is also supervised to learn *where* they are looking. Conversely, adding the label loss on top of the heatmap and angular objectives does not noticeably alter localization performance.

Nonetheless, we argue that semantic supervision remains valuable. The task of determining *what* a person is looking at, once the field of view is established, is fundamentally semantic in nature: objects within that vicinity are first identified, and context determines which of them is the most plausible target. Only after this semantic choice is made is the final spatial location defined, typically as the center of the selected object. While current benchmarks and metrics may not reveal this effect, richer evaluation settings (e.g. those involving ambiguous targets or

Method	Localization	Recognition	
	GazeAcc ↑	Acc@1 ↑	Acc@3 ↑
Random	0.166	0.002	0.006
Bias / Majority	0.352	0.082	0.083
Baseline (Mask)	0.723	0.197	0.298
Baseline (Blur)	0.723	0.306	0.458
Baseline (Crop)	0.723	0.388	0.546
Baseline (Crop FT)	0.723	<u>0.617</u>	<u>0.707</u>
Baseline (Hm weight FT)	0.723	**0.646**	**0.748**
SemGaze†	0.652	0.306	0.463
SemGaze	**0.723**	0.583	0.706

Table 8.2: Results on the GazeHOI dataset. The best scores are highlighted in bold and the second-best are underlined. Models marked with † were trained on GazeFollow and evaluated on GazeHOI without fine-tuning.

"stuff" categories) are likely to benefit from the added semantic signal.

Number of People. We observe that increasing the number of people per image during training leads to a gradual drop in performance (*cf.* Table 8.5). This effect stems from our late-fusion design: with larger N_p, the decoder must simultaneously process multiple gaze tokens, each corresponding to a different person. Consequently, the image tokens are forced to update in a way that balances and selectively aligns with several gaze queries at once, spread across different spatial regions. This makes optimization more difficult and, at the same time, places higher capacity demands on the decoder, which must disentangle and align several queries within a fixed representational budget.

Gaze Encoder Input. To assess the contribution of our gaze encoder, we test simplified variants: (i) one that uses only the head crop without positional input, and (ii) one that uses only the head box coordinates, similar to the segment-anything architecture we draw inspiration from. To compensate for the lost resolution of the head crop, this bounding box-only variant is conducted at an image resolution of 448×448 instead of 224×224.

When using only box coordinates, the burden shifts entirely to the image encoder, which must not only identify salient objects but also infer the gaze direction such that querying with a location alone retrieves the correct target. Conversely, using only the head crop likely forces the model to implicitly recover the person's position by matching head features back to the image features. As shown in Table 8.6, both alternatives lead to a significant drop in localization accuracy, confirming the necessity of combining head appearance and position for robust gaze encoding.

Gaze Decoder Depth. We ablate the number of transformer blocks used in the gaze decoder. As shown in Table 8.7, performance peaks at $N = 2$, though the differences across settings are relatively minor. This suggests that two blocks strike a good balance between capacity and

Batch Size	Avg. D. ↓	Acc@1 ↑	Acc@3 ↑
72	0.109	0.405	0.601
144	0.109	0.437	0.623
216	0.108	0.450	0.636
300	0.108	0.447	0.642

Table 8.3: Batch size ablation. Performance improves with larger batches and plateaus near 300.

Losses	Avg. D. ↓	Acc@1 ↑	Acc@3 ↑
$\mathscr{L}_{hm}$	0.107	–	–
$\mathscr{L}_{lab}$	–	0.417	0.604
$\mathscr{L}_{hm} + \mathscr{L}_{lab}$	0.108	0.451	0.640
$\mathscr{L}_{hm} + \mathscr{L}_{lab} + \mathscr{L}_{ang}$	0.108	0.447	0.642

Table 8.4: Loss composition ablation. Adding label supervision helps recognition; localization remains stable.

# People	Avg. D. ↓	Acc@1 ↑	Acc@3 ↑
$N_p = 5$	0.113	0.437	0.619
$N_p = 3$	0.112	0.433	0.620
$N_p = 1$	0.108	0.450	0.636

Table 8.5: Number of people during training. Performance degrades as N_p increases.

Gaze Encoder Inputs	Avg. D. ↓	Acc@1 ↑	Acc@3 ↑
bbox only	0.132	0.353	0.562
head crop only	0.133	0.357	0.565
bbox & head crop	0.107	0.386	0.592

Table 8.6: Gaze encoder ablation. Combining appearance (crop) and position (bbox) is crucial.

Decoder Blocks	Avg. D. ↓	Acc@1 ↑	Acc@3 ↑
$N = 1$	0.109	0.444	0.647
$N = 2$	0.108	0.447	0.642
$N = 4$	0.107	0.447	0.634

Table 8.7: Gaze decoder depth. $N = 2$ balances capacity and efficiency; deeper brings marginal gains.

Vision Pretraining	Avg. D. ↓	Acc@1 ↑	Acc@3 ↑
CLIP Vision	0.148	0.341	0.485
Supervised (ImageNet)	0.129	0.437	0.628
MAE	0.112	0.442	0.621
DINOv2	0.113	0.465	0.658
MultiMAE (ours)	0.108	0.450	0.636

Table 8.8: Image encoder pretraining. Our model is robust across inits; CLIP Vision underperforms for localization.

Text Pretraining	Avg. D. ↓	Acc@1 ↑	Acc@3 ↑
Random	0.107	0.459	0.630
BERT	0.109	0.449	0.632
MPNet	0.110	0.422	0.615
CLIP Text (FT proj)	0.108	0.439	0.633
CLIP Text (Frozen)	0.108	0.447	0.642

Table 8.9: Text pretraining. Closed-set performance is insensitive, but frozen CLIP matters for top-k and transfer.

Figure 8.5: Ablation results across architectural choices and training hyperparameters. Each mini-table reports Avg. Dist. and Acc@1/Acc@3.

efficiency, while deeper decoders provide little additional benefit.

Vision Encoder Pretraining. We evaluate the impact of different pretrained weights for the image encoder on overall model performance. In addition to our MultiMAE initialization, we test CLIP's vision encoder, a ViT trained on ImageNet, a standard MAE He et al. (2022), and DINOv2 Oquab et al. (2023). Note that DINOv2 operates at an image resolution of 224×224 with a patch size of 14, compared to 256×256 and a patch size of 16 in our default setting, though the number of output tokens remains unchanged. For memory reasons, we also reduce the batch size to 216 instead of 300.

Results are reported in Table 8.8. Using CLIP's vision weights seems to degrade performance the most (i.e. 0.148 vs. 0.108 in Avg. Dist.). By contrast, MAE and DINOv2 perform on par with MultiMAE, indicating that our architecture is relatively robust to the choice of image encoder

pretraining.

Text Encoder Pretraining. We ablate the influence of different text embeddings on model performance (*cf.* Table 8.9, second row). We compare random projections, BERT (Devlin et al., 2018), MPNet (Song et al., 2020), CLIP's text encoder with a fine-tuned projection layer, and our default setting with a frozen CLIP encoder. Somewhat surprisingly, both localization and recognition performance appear relatively insensitive to the choice of text encoder, with even random projections performing on par with pretrained models in the closed-vocabulary setting.

This result can be explained by the fact that text embeddings in our setup act as fixed *class proxies*: the model learns to align predicted embeddings to whichever vectors are provided, regardless of whether those vectors are semantically meaningful. As long as the training and evaluation vocabularies coincide, even random vectors suffice for top-1 classification. However, random embeddings fail to capture semantic structure, which becomes evident in two ways. First, they limit top-k retrieval: related classes such as *sea* and *water* do not cluster together, so Acc@3 degrades even when Acc@1 is competitive. Second, they break cross-dataset generalization: in a transfer evaluation on GazeHOI, random embeddings drop to 26% Acc@1 and 39% Acc@3, compared to 31% and 46% with frozen CLIP text embeddings.

These findings highlight that while random embeddings may appear competitive in closed-set top-1 discrimination, the use of a pretrained, semantically structured text encoder such as CLIP is essential for robust top-k performance and open-vocabulary generalization.

8.5.4 Scene Encoding as Saliency Representation

We hypothesized earlier that the architectural separation between scene and gaze encoding should encourage the image encoder to highlight potential gaze target candidates in a person-agnostic manner. To examine this, we visualize in the second row of Figure 8.6 the attention maps from the final layer of the image encoder. The visualizations confirm our expectation: the encoder consistently emphasizes not only salient regions but also objects and areas that are likely gaze targets (e.g. heads, hands, documents, tables, signboards), indicating that it constructs a candidate set of gaze-relevant locations before any person-specific prompting is applied. This supports our interpretation of the decoder as a filtering mechanism that selects the correct target from among these candidates.

8.5.5 Role of the Background Token

Quantitatively, the addition of the background token improves performance, particularly on localization metrics (e.g. Avg. Dist. of 0.112 vs. 0.115 for models trained with $N_p = 3$ and a batch size of 216). To better understand this effect, we visualize the *scene-to-gaze* ($\mathbf{x}_s \rightarrow \mathbf{x}_g$) cross-attention maps from the final decoder block. With a single input person ($N_p = 1$), each

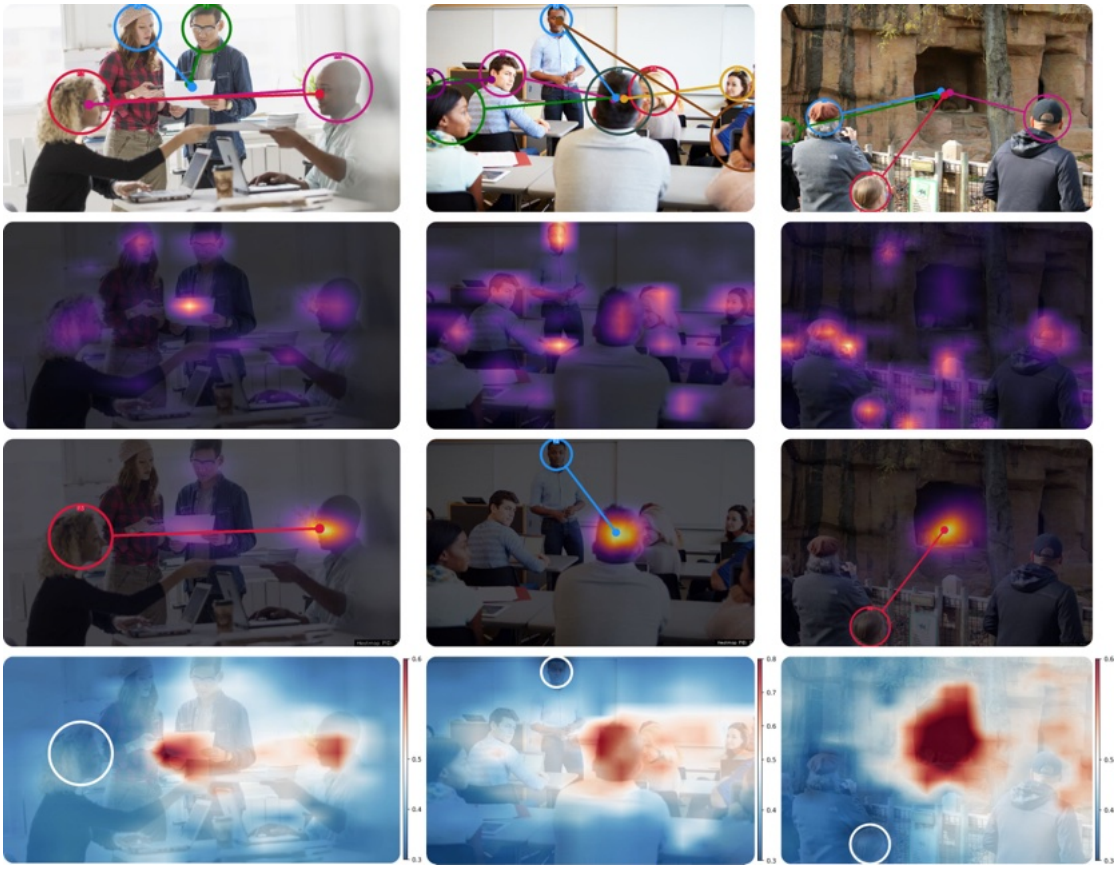


Figure 8.6: Qualitative results of our model on internet images. **Top row:** predicted gaze points for all detected people. **Second row:** attention maps from the final layer of the image encoder, highlighting salient regions. **Third row:** predicted gaze heatmap for a selected person. **Bottom row:** weights from the final image-to-gaze cross-attention layer of the decoder for the same person, showing how gaze and background tokens are distributed across the scene.

image token is updated as a weighted average of the gaze token and the background token. The last row of Figure 8.6 shows the gaze-token weight w (with the background weight given by $1 - w$): red regions correspond to image tokens attending more strongly to the gaze token, while blue regions correspond to those dominated by the background token.

These visualizations confirm that the decoder leverages the background token as a placeholder for unattended regions, effectively separating gaze-relevant areas from irrelevant context. This mechanism enhances the model's ability to discriminate the true gaze target and contributes to the observed performance gains.

8.5.6 Revisiting Efficiency

We previously noted that our baseline carries a higher parameter count than the proposed model. Here, we provide a FLOPs comparison as a function of the number of people in the image to illustrate how our architecture scales. Alongside our model and baseline, we also report the results of the efficiency study from Chapter 6, including Sharingan (Tafasca et al.,

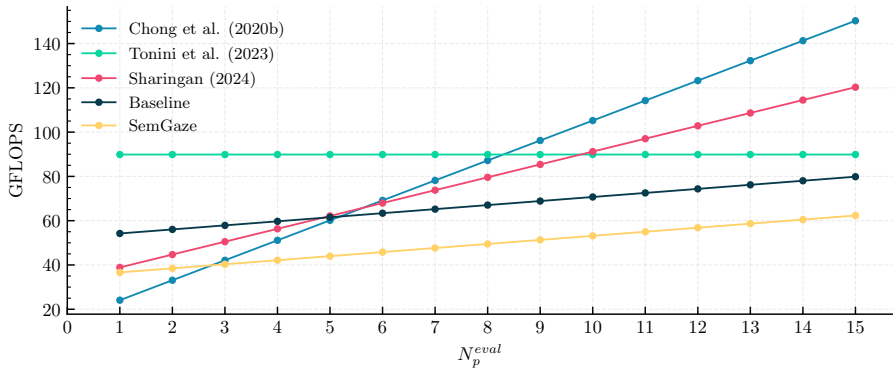


Figure 8.7: Comparative analysis of FLOPS vs number of people during inference.

2024b).

Figure 8.7 shows that the one-stage model of Tonini et al. (2023) has constant cost, while the early-fusion approach of Chong et al. (2020b) grows rapidly, since the expensive scene encoding is repeated for each person. By contrast, our model exhibits only a modest increase thanks to its lightweight decoder and the fact that the scene encoding is performed once per image. The baseline is consistently heavier by a constant 20 GFLOPs due to the additional ViT backbone, which is only run once per image.

Importantly, SemGaze is also more efficient than our previous Sharingan architecture (Tafasca et al., 2024b). While Sharingan already reduced cost compared to early-fusion methods, its joint processing of scene and gaze tokens meant that FLOPs still scaled more sharply with N_p. By decoupling the two streams and adopting a promptable decoder, SemGaze achieves noticeably better efficiency and remains significantly more efficient than the constant-cost model of Tonini et al. (2023) even at $N_p = 15$.

8.5.7 Qualitative Evaluation

We present qualitative results on the GazeHOI dataset, including both correct predictions and failure cases in Figure 8.8.

Overall, common categories such as *book*, *cellphone*, and *ball* are typically well localized and correctly identified. The model performs robustly across diverse conditions, including variations in scene type, number of people, activity, interaction complexity, distance from the camera, viewing angle, and lighting.

Failure cases fall into several categories. First, some are inherently difficult because the object is barely visible (i.e. either small, occluded, or partially cropped). In these cases, the model tends to predict the nearest larger object in its vocabulary. For instance, in the second image of the last row, the target *bag* is hard to spot, leading the model to predict the nearby animal (mapped to *giraffe* or *zebra*, since *deer* is not in the vocabulary). Second, errors can occur when the ground-truth label is correct but the location is off, often in images with multiple

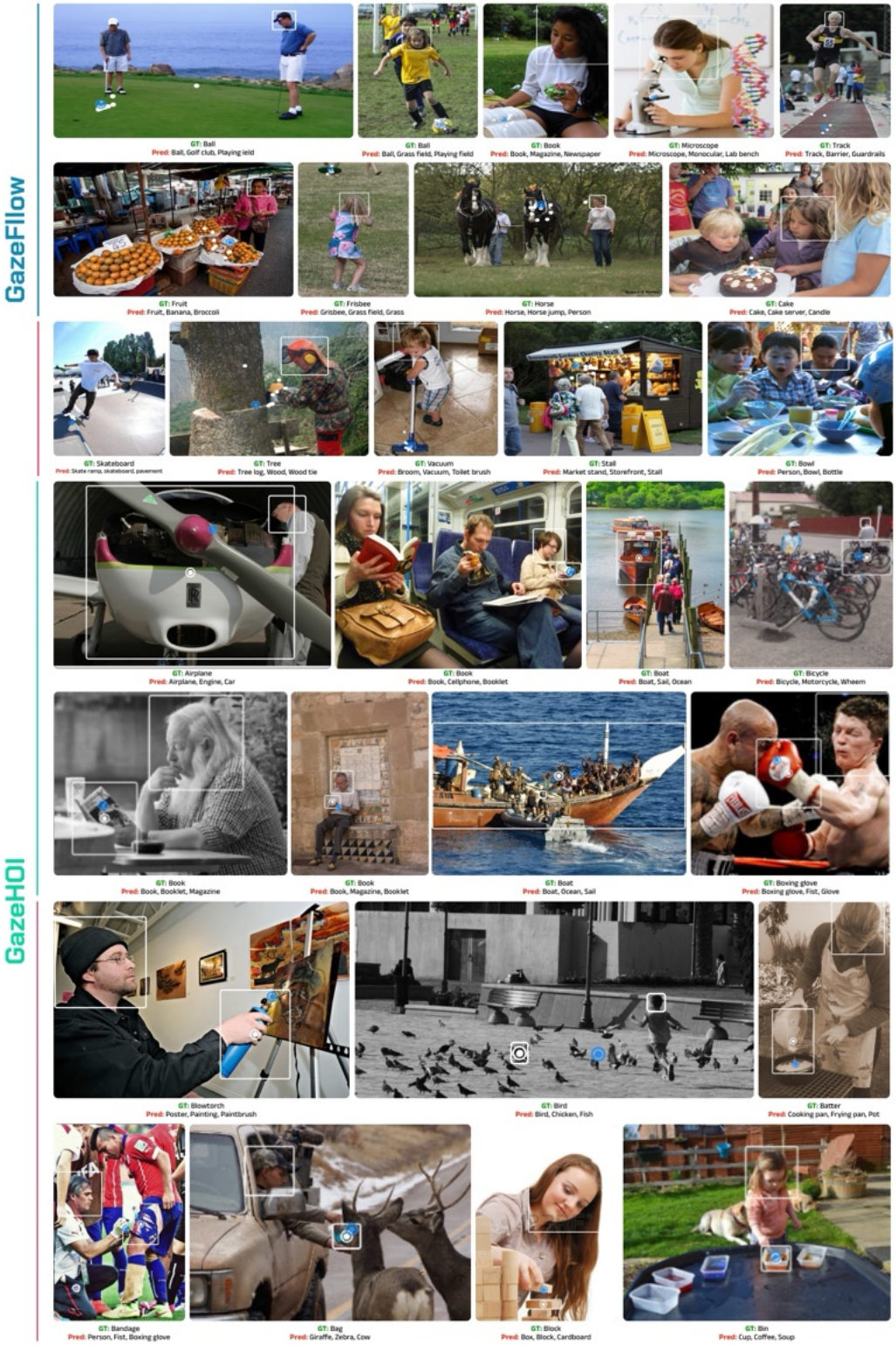


Figure 8.8: Qualitative examples from GazeFollow and GazeHOI. Ground-truth annotations are shown in white and predictions in blue. For each sample, we also display the top-3 predicted class labels alongside the ground-truth label. Rows highlighted with a pink line on the left indicate failure cases, where either the predicted class label or the gaze location does not match the ground truth.

instances of the same class (e.g. predicting the wrong *bird* in the second image of the sixth row). Finally, some failures reflect the multi-label nature of the task: in the first image of the sixth row, the model predicts *poster* and *painting*, while the ground-truth is *blowtorch*. Both *painting* and *blowtorch* are indeed in the person's field of view, but the dataset annotation specifies only one.

8.6 Limitations

While our model demonstrates strong performance in recognizing gaze target classes, the current design has several limitations. First, it is not equipped to handle the inherently multi-label nature of gaze. In many situations, the gaze heatmap may overlap with multiple objects, or the target may be semantically ambiguous. Ideally, the model should be able to output several plausible classes (e.g. *knife*, *onion*, and *chopping board*), rather than a single label embedding. Since our architecture collapses all supervision into one prediction, it risks an "averaging effect" over possible object embeddings, reminiscent of the blurring observed in Chapter 6 when regressing gaze coordinates directly instead of predicting a heatmap.

Second, while the model produces both a gaze location (heatmap) and a gaze label embedding, these two predictions are optimized with separate supervision and are not explicitly forced to match. This can create inconsistencies: the model may localize correctly but misclassify the object, or predict the correct class while the heatmap is shifted to a nearby region. This is not unlike object detection, where localization and recognition can also disagree; however, unlike detection, gaze does not currently have a well-established object-centric evaluation protocol to reconcile these cases. Ultimately, we argue that a more principled, object-centric evaluation is needed, similar to the setup in GazeHOI which contains bounding boxes of targets. A prediction can then be considered correct if the gaze point falls within the ground-truth bounding box and the predicted label matches the ground-truth class. This would provide a clearer definition of true positives and negatives, and could naturally extend to multi-label settings.

Finally, dataset scale and supervision remain a bottleneck. GazeFollow, for instance, contains many more objects in each scene than annotated gaze instances. These unlabelled objects represent missed opportunities for supervision. Leveraging them through auxiliary grounding losses (e.g. aligning spatially localized features with textual labels, as explored in Li et al. (2022b)), could enhance recognition and improve the robustness of the learned gaze embeddings.

8.7 Conclusion

In this chapter, we introduced SemGaze, a promptable transformer architecture that jointly predicts where a person is looking and what they are looking at. By decoupling scene and gaze encoding, SemGaze achieves both state-of-the-art localization performance and competitive

recognition accuracy, while remaining more efficient and scalable than prior designs such as Sharingan. Extensive ablations confirmed the value of semantic supervision, the role of background tokens in filtering, and the robustness of our approach across different pretraining strategies. At the same time, we highlighted key limitations: the inability to capture multi-label gaze, inconsistencies between localization and recognition outputs, and the lack of object-centric evaluation protocols.

These findings suggest that semantic grounding provides a critical step toward richer and more interpretable gaze understanding, but also underline the need for output representations that go beyond discrete points or heatmaps. In particular, while heatmaps model multimodality and spatial uncertainty, they remain non-parametric and difficult to interpret quantitatively beyond an arg max aggregation. The next chapter therefore turns to predictive uncertainty, introducing a probabilistic framework that treats gaze prediction as a parametric probability distribution rather than a single location. This perspective enables a principled treatment of multimodality and variance, and further allows each mode of the distribution to be associated with its own semantic class embedding, which offers an elegant solution to the semantic averaging effects observed in the current design.

9 Predictive Uncertainty

Images, people, and semantics across the previous chapters all conspired to make gaze prediction *ambiguous*. Geometry (Chapter 5) showed that many 3D targets project to the same 2D ray; multi-person scenes (Chapter 6) added competing, socially meaningful candidates; social interpretation (Chapter 7) made correctness relational and context-dependent; and semantic grounding (Chapter 8) required mapping gaze to discrete entities, exposing multimodality and the limitations of single-embedding outputs. This chapter, based on unpublished work, argues that handling such ambiguity cannot be an afterthought of the decoder or merely a property of the evaluation metric. It must be a *first-class representation*: instead of predicting a point or a heatmap, we predict a *distribution* over gaze targets (*cf.* Figure 9.1).

We build on the promptable architecture of Chapter 8 while reintroducing the 3D Field-of-View (FoV) mechanism from Chapter 5 to provide geometric conditioning. In this framework, the spatial decoder is replaced with a lightweight MLP that predicts the parameters of a mixture distribution over 2D gaze locations, trained under a maximum-likelihood objective. This modification is not a superficial swap of heads: by predicting a distribution, the model can *express* uncertainty rather than merely *suffer* from it, while the 3D FoV supplies a geometric prior that constrains physically implausible targets. Together, these additions enable both geometric grounding and probabilistic reasoning across multiple plausible gaze hypotheses. The following section motivates this shift and formalizes its benefits.

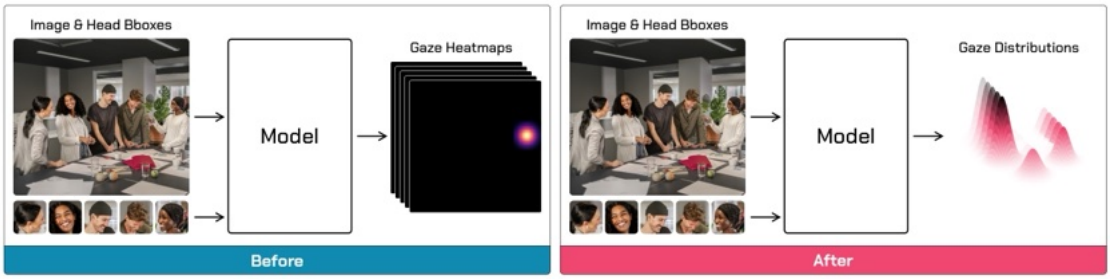


Figure 9.1: Illustration of the difference between previous methods that produce non-parametric gaze heatmaps, and our new method that outputs a probability distribution for the gaze location.

9.1 Motivation

From Heatmaps to Likelihoods. Heatmaps extended point regression by allowing multiple spatial modes, but they remained heuristic: dense, discretized confidence maps trained with pixel-wise regression losses. Their values express activation rather than calibrated probability, and recovering explicit hypotheses still requires post-hoc peak finding. In this chapter, we replace these implicit confidence maps with a likelihood-based formulation that directly models $p(\mathbf{y}|\mathbf{x})$ as a continuous, normalized probability density. This shift grounds gaze prediction in probabilistic reasoning, where each mode is a hypothesis weighted by its likelihood, providing a compact, interpretable, and statistically principled representation of uncertainty.

Uncertainty as Signal. Natural scenes often contain multiple plausible gaze targets. Ambiguity arises from *aleatoric factors* inherent to the scene (e.g. occlusions, partial visibility of the eyes, or image quality), and from *epistemic factors* stemming from model or data limitations. Even with single-point annotations, the underlying perceptual distribution of gaze targets may be multimodal, reflecting several possible outcomes with different contextual likelihoods. A useful model should capture this structure: be confident and decisive when the scene is unambiguous, and explicitly multimodal when it is not. Figure 9.2 illustrates the distinction between aleatoric and epistemic uncertainty in the context of gaze target prediction.

9.1.1 Advantages of a Probabilistic Formulation

Compact Output and Efficient Decoding. Instead of relying on large deconvolutional or spatial decoders to generate full-resolution heatmaps, the model directly predicts a compact set of distribution parameters (means, covariances, and mixing weights) through a simple MLP. This greatly reduces computation, memory usage, and training time.

Resolution-Independent Representation. The probabilistic output is spatially continuous and independent of the heatmap resolution, avoiding discretization artifacts and enabling consistent evaluation at any image scale.

Direct Access to Multiple Modes. The mixture formulation explicitly models multimodality: each Gaussian component can correspond to a plausible gaze hypothesis. Unlike heatmaps, which require heuristic peak detection to extract multiple modes, the mixture provides them in closed form through the predicted parameters.

Explicit and Quantitative Uncertainty. Each component encodes its own variance (or covariance), providing a natural measure of confidence for every hypothesis. This fine-grained uncertainty is difficult to obtain from heatmaps, whose spread is often arbitrary and resolution-dependent.

Flexible Decision Strategies. Because the output is a full probability distribution, the model



Figure 9.2: Illustration of *epistemic* and *aleatoric* uncertainty in gaze following. **Left:** the model fails because it has likely never seen top-down, bird's-eye-view scenes during training. This uncertainty stems from limited data coverage and can be reduced by exposing the model to more diverse viewpoints. This is an example of *epistemic* uncertainty. **Right:** the subject's face is occluded, leaving insufficient visual evidence to determine whom they are looking at. Here, the ambiguity is intrinsic to the scene and cannot be resolved even with more data or a better model. This is an example of *aleatoric* uncertainty.

allows multiple inference modes: i) selecting the most probable component (highest mixing weight), ii) the most certain one (lowest variance), iii) or combining both criteria (highest expected likelihood). It can also aggregate several modes when downstream tasks benefit from broader coverage or multi-target reasoning.

Statistical Soundness and Unified Training Objective. The formulation directly optimizes the likelihood of the ground-truth point under the predicted distribution, ensuring a principled statistical objective. It also naturally extends to multi-annotator data, where each annotation can be treated as a sample from the true underlying distribution.

9.1.2 From Point Supervision to Distributional Prediction

Existing gaze following datasets provide point-level supervision: for each person in an image, annotators select a single pixel location corresponding to where the person is looking. While simple to collect and evaluate, this design carries an implicit but critical assumption about the annotation process and the nature of uncertainty in gaze data.

Annotation assumption. Two conceptual models of the annotation process can be considered:

Under a *mode-selection assumption*, the annotator consistently selects the most likely target (i.e. the mode of the underlying distribution). This preserves the relative likelihoods between competing hypotheses: if two objects are plausible but one is contextually more salient (e.g. a speaking person), annotations will predominantly favor that mode. The resulting dataset encodes structured ambiguity, where multimodality exists but with uneven probability mass reflecting true contextual biases.

Under a *random-sampling assumption*, the annotation corresponds to a random draw from the set of plausible targets, regardless of their likelihood. Across many samples, this would flatten the true distribution, assigning approximately equal probability to all modes even when one is more probable. Such supervision injects uninformative entropy, obscuring the structure of gaze uncertainty rather than capturing it.

In practice, the *mode-selection* assumption is the more realistic one for gaze following datasets. Human annotators are not asked to sample randomly among plausible targets but to identify the location that best reflects where the person appears to be looking. When several candidates exist, they naturally select the most convincing hypothesis based on contextual cues such as saliency, social signals, or object relevance. It is therefore reasonable to treat annotations as *mode-seeking* rather than *stochastic* samples, and to design training objectives that reflect this behavior.

Alignment with maximum likelihood training. The maximum likelihood objective used in probabilistic modeling aligns naturally with the *mode-selection* assumption. When each annotation represents the most likely target, maximizing the probability of the observed point encourages the model to concentrate probability mass around the dominant mode while maintaining smaller secondary peaks when the scene is ambiguous. This yields asymmetric, contextually calibrated distributions: sharp and decisive when the gaze is unambiguous, and multimodal yet weighted when multiple hypotheses coexist.

In contrast, if the annotations were random samples, the same objective would lead the model to reproduce uniformly distributed modes, effectively learning to mirror annotation noise rather than genuine perceptual uncertainty. Hence, the *mode-selection* assumption ensures that probabilistic training reflects meaningful, context-grounded uncertainty rather than artificial randomness.

Implications for evaluation. This perspective also clarifies why standard point-based evaluation metrics, such as minimum distance or AUC, remain valid despite multimodality. These metrics implicitly reward accurate localization of the dominant mode (i.e. the one selected by annotators), while tolerating probability mass elsewhere. Probabilistic modeling thus provides a principled framework that accommodates both the deterministic nature of point supervision and the inherent uncertainty of real-world gaze behavior.

In summary, this chapter motivates and develops a probabilistic gaze target prediction framework named *ProbGaze*. We (i) review related work on uncertainty in deep learning; (ii) detail the probabilistic decoder and training objective; (iii) show that mixture outputs achieve nearly the same accuracy as heatmap-based predictions with a lighter head while exposing usable uncertainty; (iv) analyze uncertainty quality via uncertainty–error correlation; and (v) demonstrate how per-mode semantics resolve the “single-embedding” limitation of Chapter 8. We close by reflecting on the broader implications of predictive uncertainty and how it connects to the themes of the subsequent chapters.

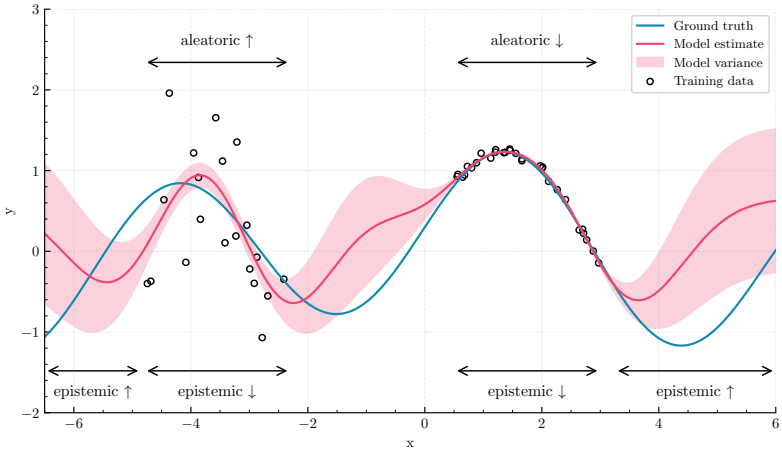


Figure 9.3: Illustration of the difference between aleatoric and epistemic uncertainty. Aleatoric uncertainty increases where observations are noisy or inconsistent, while epistemic uncertainty grows outside the training data manifold, where the model must extrapolate.

9.2 Related Work

9.2.1 Uncertainty in Deep Learning

Taxonomy and scope. Uncertainty in deep learning is commonly partitioned into aleatoric and epistemic components. *Aleatoric uncertainty* captures the inherent ambiguity or noise in the data itself—arising from factors such as occlusions, motion blur, low image quality, or the presence of multiple plausible outcomes that cannot be disambiguated from the evidence alone. *Epistemic uncertainty*, in contrast, reflects the model's lack of knowledge: it is high in regions where training data is sparse or out-of-distribution, and it can in principle be reduced by collecting more data or improving the model's capacity (*cf.* Figure 9.3). Kendall and Gal (2017) popularized this split for vision systems and showed how to combine both within a single framework for segmentation and depth tasks. Recently, some works (Hüllermeier and Waegeman, 2021; Smith et al., 2024) caution that the terminology is used inconsistently and advocate tying definitions to predictive distributions and information gain; we adopt the terms but are explicit about what we measure. Our chapter focuses on predictive/aleatoric uncertainty (i.e. the inherent ambiguity of where someone might be looking) by explicitly parameterizing $p(\mathbf{y}|\mathbf{x})$.

How is uncertainty estimated? Three main families of approaches dominate uncertainty estimation in deep learning. (i) *Bayesian or variational approximations*, such as Monte Carlo Dropout (Gal and Ghahramani, 2016), approximate Bayesian inference by treating dropout at test time as posterior sampling over network weights, thereby capturing epistemic uncertainty. (ii) *Deep ensembles* (Lakshminarayanan et al., 2017) estimate epistemic uncertainty through diversity across independently trained models, offering strong empirical calibration and robustness without requiring explicit Bayesian machinery. (iii) *Heteroscedastic regression* (Le et al., 2005; Kendall and Gal, 2017) learns input-dependent observation noise (e.g. variance) to capture aleatoric uncertainty directly within the model's likelihood, enabling the network

to express confidence about inherently ambiguous or noisy data. In this chapter, we focus on the third family: modeling the output distribution itself rather than uncertainty over model weights, since gaze following naturally involves intrinsic scene ambiguity (multiple plausible gaze targets, occlusions, or limited visual evidence).

Evidence across vision tasks. Uncertainty has improved reliability in a range of core vision problems: semantic segmentation (Kendall et al., 2015), depth/segmentation with learned attenuation (Kendall and Gal, 2017), camera relocalization and pose regression where uncertainty guides the balance between translation/rotation (Kendall and Cipolla, 2017), and object detection where Gaussian heads model localization variance (Choi et al., 2019). Generative formulations like the Probabilistic U-Net (Kohl et al., 2018) further show how to represent multiple plausible outputs in inherently ambiguous settings. These strands motivate treating uncertainty as a first-class output rather than post-hoc scores.

Loss shaping for heteroscedastic regression. Maximizing the Gaussian Negative Log-Likelihood (NLL) for (μ, σ) can under-train the mean when the variance is large, since the squared error is down-weighted by the variance. Seitzer et al. (2022) analyze this pitfall and propose β-NLL, which scales each sample's loss by its (detached) σ^β, restoring gradient pressure on hard, high-uncertainty examples and improving both Root Mean Square Error (RMSE) and log-likelihood across domains. A follow-up work from Stirn et al. (2023) takes the idea further, ensuring heteroscedastic models match the mean accuracy of standard regression by canceling variance effects in the gradient and decoupling variance updates from shared features. This diagnosis (i.e. variance attenuates supervision) is directly relevant to our setting which uses a Gaussian Mixture Model (GMM) instead of a simple Gaussian. In our case, we address the problem by introducing a *winner-takes-all* auxiliary loss term that sharpens the mean of the "responsible" hypothesis without discarding uncertainty.

Uncertainty in gaze. Uncertainty estimation has also been explored in gaze-related tasks. The model of Kellnhofer et al. (2019) predicts both a 3D gaze vector and its associated uncertainty, trained with a Pinball loss to improve robustness and generalization. More recently, Miao et al. (2025) incorporated uncertainty estimates into 3D gaze vector prediction to fuse information across camera views. By contrast, no prior work has modeled an explicit spatial distribution over 2D gaze targets in the gaze following setting. This chapter fills that gap by introducing a compact, parametric output that is both interpretable and computation-friendly.

9.2.2 Distributional and Multi-Modal Regression

Mixture Density Networks. A natural way to move from point prediction to full uncertainty modeling is to represent the conditional target distribution $p(\mathbf{y}|\mathbf{x})$ explicitly through a parametric mixture model. This idea was first formalized in the Mixture Density Network (MDN) framework (Bishop, 1994), where a neural network predicts the parameters of a Gaussian mixture (i.e. means, variances, and mixing coefficients), instead of a single deterministic

output. MDNs have since become a standard tool for multi-modal regression problems where multiple outcomes can plausibly arise from the same input, including human pose and motion forecasting (Li and Lee, 2019), trajectory prediction (Makansi et al., 2019), and object detection (Varamesh and Tuytelaars, 2020). In such settings, the mixture components represent distinct hypotheses, and training by maximum likelihood automatically learns their relative weighting according to data frequency. In our context, this approach naturally captures the intrinsic multi-modality of gaze following, where two or more plausible gaze targets may coexist within the same scene. Unlike heatmaps, which provide a discretized, unnormalized estimate of spatial probability, mixture models offer a continuous and compact representation with explicit per-mode uncertainty. That being said, MDNs are notoriously difficult to train in practice, often suffering from instability and mode collapse, where multiple components converge to similar means or a single dominant mode (Rupprecht et al., 2017).

Multiple Hypothesis Prediction. An alternative but closely related direction is Multiple Hypothesis Prediction (MHP) (Rupprecht et al., 2017), which approximates multi-modal distributions through a discrete set of parallel predictions. Instead of learning soft responsibilities as in an MDN, MHP employs a winner-takes-all (WTA) loss that updates only the hypothesis closest to the ground truth. Each output head thus specializes in a different mode of the target space, leading to a piecewise approximation of $p(\mathbf{y}|\mathbf{x})$ without requiring explicit mixture weights or variances. This formulation has been adopted in several domains, including optical flow (Ilg et al., 2018) and human pose estimation (Rommel et al., 2023), where multiple plausible outputs exist for a given observation. While effective at maintaining hypothesis diversity, MHP remains non-probabilistic: its outputs are deterministic samples rather than calibrated distributions. Our probabilistic mixture formulation can be viewed as a continuous extension of this idea: we retain the interpretability and sharpness of the WTA mechanism through an auxiliary loss, while grounding the overall prediction in a normalized likelihood that provides explicit uncertainty and mode weighting.

9.2.3 Alternative Distributional Models

Beyond Gaussian mixtures. While Gaussian mixture models (GMMs) are theoretically universal density approximators, they become inefficient in practice for high-dimensional data, as accurately modeling complex densities often requires an impractically large number of components. This limits their scalability beyond low-dimensional tasks such as gaze target prediction.

Normalizing flows. Normalizing flows (Dinh et al., 2016; Papamakarios et al., 2021) overcome this limitation by learning bijective transformations between data and a simple latent distribution, enabling exact likelihood computation and smooth interpolation in latent space. However, these models impose strict architectural constraints, where each layer must be invertible with tractable Jacobians, which restricts flexibility and increases implementation

complexity.

Diffusion models. Diffusion-based generative models (Ho et al., 2020) represent a more recent and powerful alternative, capable of modeling highly complex, multimodal densities with exceptional fidelity. Yet, their iterative sampling and training procedures make them computationally demanding and data-hungry, particularly unsuitable for tasks requiring lightweight and real-time inference.

In contrast, our probabilistic formulation adopts a low-dimensional Gaussian mixture model that strikes a balance between expressivity and tractability: sufficient to capture the multimodal nature of gaze behavior while remaining efficient and interpretable.

9.3 Architecture

Our model, *ProbGaze*, builds directly upon the promptable architecture of *SemGaze*, preserving its scene encoder, gaze encoder, and bidirectional transformer decoder. In addition, it reintroduces the concept of a 3D Field-of-View (FoV) from *GeomGaze* (Chapter 5), integrating geometric cues through FiLM-based conditioning (Perez et al., 2018). The key innovation lies in the output gaze representation: rather than decoding spatial heatmaps, *ProbGaze* predicts the parameters of a conditional probability distribution over 2D gaze coordinates, reframing gaze estimation as a probabilistic likelihood modeling task rather than deterministic heatmap regression.

9.3.1 3D Field-of-View Conditioning

To incorporate geometric constraints, we extend the decoder with a *3D Field-of-View (FoV) Encoder*, following the formulation introduced in Chapter 5. Given the predicted 3D gaze vector and the scene depth map, we first construct a 3D FoV heatmap that encodes the directional agreement between the gaze vector and the scene geometry. This FoV heatmap is then passed through a lightweight convolutional block to produce modulation parameters $\boldsymbol{\gamma}_{\text{fov}}, \boldsymbol{\beta}_{\text{fov}} \in \mathbb{R}^{N \times D}$, aligned in dimension with the scene tokens $\mathbf{x}_s$. These parameters are used to condition the scene representation via a FiLM-style operation (Perez et al., 2018) that allows the network to modulate or even disregard the FoV features when they are uninformative.

$$\begin{aligned} \mathbf{x}_s &\leftarrow \text{Norm}(\mathbf{x}_s), \\ \mathbf{x}_{s,p} &= \mathbf{x}_s \odot (1 + \boldsymbol{\gamma}_{\text{fov}}) + \boldsymbol{\beta}_{\text{fov}}, \end{aligned} \tag{9.1}$$

where Norm$(\cdot)$ denotes feature-wise normalization (Layer Normalization for token sequences or Group Normalization with group size 1 for spatial feature maps), and $\odot$ denotes element-wise multiplication. The modulated scene tokens are then processed by the gaze decoder. This design provides a flexible geometric prior that enhances spatial reasoning while main-

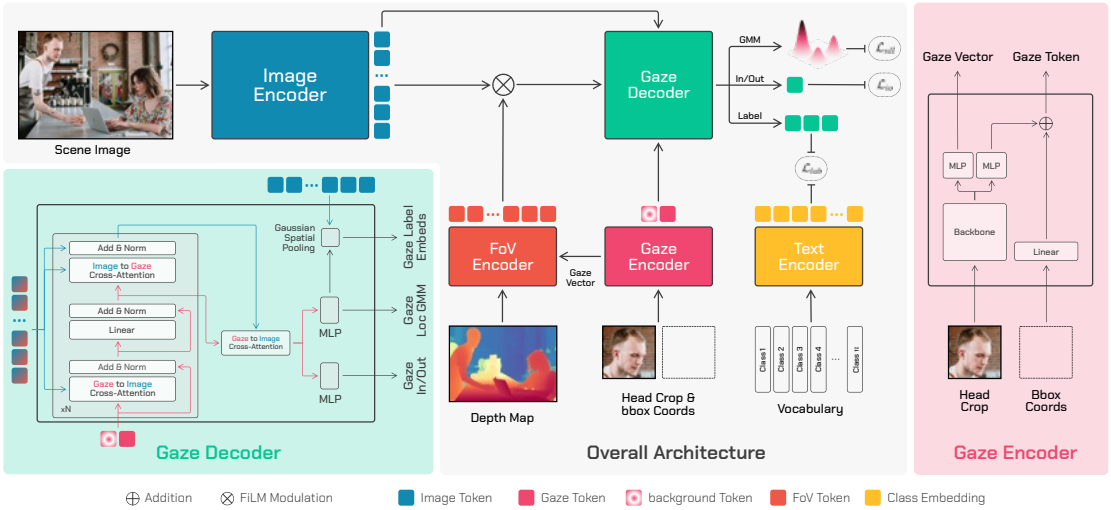


Figure 9.4: Overview of the proposed ProbGaze architecture. The backbone and promptable design are identical to SemGaze (Chapter 8), with three main modifications: (i) we leverage a 3D FoV based on extracted depth information; (ii) the decoder now predicts Gaussian Mixture Model (GMM) parameters instead of spatial heatmaps, skipping the upscaler and using the MLP head directly; and (iii) multiple label embeddings are generated per person (one for each mixture component).

taining the model's ability to rely purely on appearance cues when depth or gaze geometry are unreliable.

9.3.2 Probabilistic Decoder

The person-conditioned scene tokens $\mathbf{x}_{s,p} \in \mathbb{R}^{N \times D}$ and gaze token $\mathbf{x}_g \in \mathbb{R}^D$ are provided to the decoder following the same procedure described in Chapter 8. After the final decoder block, the updated gaze token $\mathbf{x}_g$ is mapped to a set of Gaussian Mixture Model (GMM) parameters using a dedicated MLP head:

$$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K} = \mathrm{MLP}_{\mathrm{GMM}}(\mathbf{x}_g), \tag{9.2}$$

where $\pi_k \in \mathbb{R}$ are the mixture weights, $\boldsymbol{\mu}_k \in \mathbb{R}^2$ the 2D gaze means, and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{2\times 2}$ the covariance matrices for each of the K components. This lightweight head directly predicts the GMM parameters, and skips the upscaler module used for heatmap generation in SemGaze, resulting in a more compact and resolution-independent decoder.

To ensure numerical stability and allow flexible uncertainty modeling, we handle the covariance matrix differently depending on the chosen parameterization. In the *isotropic* case, a single unconstrained scalar $\tilde{\sigma}_k \in \mathbb{R}$ is predicted for each component and mapped to a positive standard deviation using the `softplus` function:

$$\sigma_k = \mathrm{softplus}(\tilde{\sigma}_k) + \varepsilon, \qquad \boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}. \tag{9.3}$$

For the *diagonal* case, the model outputs two unconstrained values $\tilde{\boldsymbol{\sigma}}_k = [\tilde{\sigma}_{k,x}, \tilde{\sigma}_{k,y}] \in \mathbb{R}^2$ which are converted independently:

$$\boldsymbol{\sigma}_k = \text{softplus}(\tilde{\boldsymbol{\sigma}}_k) + \varepsilon, \qquad \boldsymbol{\Sigma}_k = \text{diag}(\boldsymbol{\sigma}_k^2). \tag{9.4}$$

For the *full covariance* case, we predict three unconstrained parameters $\tilde{\mathbf{l}}_k = [\tilde{l}_{k,00}, \tilde{l}_{k,10}, \tilde{l}_{k,11}] \in \mathbb{R}^3$ defining the lower-triangular Cholesky factor $\mathbf{L}_k \in \mathbb{R}^{2\times 2}$:

$$\mathbf{L}_k = \begin{bmatrix} \text{softplus}(\tilde{l}_{k,00}) + \varepsilon & 0 \\ \tilde{l}_{k,10} & \text{softplus}(\tilde{l}_{k,11}) + \varepsilon \end{bmatrix}, \qquad \boldsymbol{\Sigma}_k = \mathbf{L}_k \mathbf{L}_k^\top. \tag{9.5}$$

This parameterization guarantees that all predicted covariance matrices are positive definite while preserving unconstrained optimization in the network's output space.

9.3.3 In/Out Prediction

An additional MLP head, $\text{MLP}_{\text{inout}}$, predicts whether the gaze target lies inside or outside the image:

$$p_{\text{in}} = \sigma\left(\text{MLP}_{\text{inout}}(\mathbf{x}_g)\right), \tag{9.6}$$

where σ denotes the sigmoid activation.

9.3.4 Per-Mode Semantic Embeddings

Each Gaussian component $(\boldsymbol{\mu}_k, \sigma_k)$ defines a distinct hypothesis about the attended location in the scene. To assign a semantic interpretation to each hypothesis, we derive a mode-specific semantic embedding directly from the image feature map $\mathbf{x}_s \in \mathbb{R}^{H\times W\times D}$ produced by the frozen backbone (before FoV fusion). Unlike previous variants that condition on the gaze token, this formulation lets each mixture mode query the scene independently, integrating only the visual evidence lying within its predicted field of attention.

Spatial Gaussian Attention. Let $\mathbf{p}_{u,v} \in [0,1]^2$ denote the normalized coordinates of position (u, v) on the feature grid. For each mode k, we define a geometric attention prior that measures the proximity of each spatial position to the predicted mean:

$$b_k(u, v) = -\tfrac{1}{2} \frac{\|\mathbf{p}_{u,v} - \boldsymbol{\mu}_k\|_2^2}{\sigma_k^2}. \tag{9.7}$$

A learnable global scale parameter α controls the overall sharpness of this distribution:

$$\ell_k(u,v) \;=\; \alpha\, b_k(u,v), \tag{9.8}$$

so that larger values of α produce narrower, more concentrated weighting around $\boldsymbol{\mu}_k$, while smaller values broaden it. Unlike cross-attention mechanisms, the weights depend solely on the predicted spatial parameters and do not incorporate any content-based similarity term; the operation therefore acts as a purely geometric, probabilistic pooling of frozen visual features.

Normalized Attention Weights. The unnormalized logits $\ell_k(u,v)$ are converted into attention weights through a spatial softmax:

$$a_k(u,v) = \frac{\exp\big(\ell_k(u,v)\big)}{\sum_{u',v'} \exp\big(\ell_k(u',v')\big)}. \tag{9.9}$$

This results in a continuous and differentiable distribution over the feature map, centered at $\boldsymbol{\mu}_k$ and scaled by σ_k. Intuitively, σ_k controls the relative spatial uncertainty the model attributes to each gaze hypothesis, while α calibrates this uncertainty globally across all samples, effectively adjusting the receptive field for the pooling operation.

Context extraction. The local context for each mode is then obtained as a weighted average of the image features:

$$\mathbf{c}_k = \sum_{u,v} a_k(u,v)\,\mathbf{x}_s(u,v), \quad \mathbf{c}_k \in \mathbb{R}^D. \tag{9.10}$$

Semantic Projection. A lightweight readout head maps this context to a CLIP-compatible semantic vector:

$$\mathbf{x}_{\ell,k} = \mathrm{MLP}_{\mathrm{label}}(\mathbf{c}_k), \quad \mathbf{x}_{\ell,k} \in \mathbb{R}^{D_\ell}, \tag{9.11}$$

yielding one semantic embedding per mixture component.

In summary, ProbGaze reuses the full SemGaze architecture but introduces probabilistic output modeling, per-mode semantics, and an explicit in/out prediction head, yielding a compact, interpretable, and uncertainty-aware formulation of gaze following.

9.4 Implementation Details

Loss Function. We optimize the model with a weighted sum of four terms:

$$\mathscr{L} = \lambda_{\mathrm{nll}}\,\mathscr{L}_{\mathrm{nll}} + \lambda_{\mathrm{dist}}\,\mathscr{L}_{\mathrm{dist}} + \lambda_{\mathrm{ang}}\,\mathscr{L}_{\mathrm{ang}} + \lambda_{\mathrm{lab}}\,\mathscr{L}_{\mathrm{lab}}. \tag{9.12}$$

(i) Mixture NLL. Given the ground-truth 2D gaze point $\mathbf{y} \in \mathbb{R}^2$ and the predicted Gaussian

Mixture Model (GMM) $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$, we define:

$$\mathscr{L}_{\text{nll}} = -\log\left(\sum_{k=1}^{K} \pi_k \, \mathcal{N}\big(\mathbf{y} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\big)\right). \tag{9.13}$$

(ii) Distance loss. Let the winning component be the one whose mean is closest to the ground truth:

$$k^\star = \arg\min_k \|\mathbf{y} - \boldsymbol{\mu}_k\|_2, \qquad \mathscr{L}_{\text{dist}} = \|\mathbf{y} - \boldsymbol{\mu}_{k^\star}\|_2^2. \tag{9.14}$$

This auxiliary term addresses a known limitation of heteroscedastic and mixture likelihood training, where the variance term in the denominator of the loss can down-weight challenging or ambiguous samples, leading to under-optimized mean estimates. Similar effects were analyzed in Seitzer et al. (2022) and later revisited by Stirn et al. (2023), who proposed balancing or decoupling the mean and variance gradients. Our *winner-takes-all* formulation provides an equivalent rebalancing mechanism by explicitly sharpening supervision on the component that best explains each sample, thereby preserving localization accuracy without disrupting probabilistic calibration.

(iii) Angular loss. Identical to SemGaze, we include an angular regression loss on the normalized gaze vector $\mathbf{g}_v \in \mathbb{R}^3$ produced by the gaze encoder:

$$\mathscr{L}_{\text{ang}} = 1 - \frac{\langle \mathbf{g}_v, \mathbf{g}_v^{gt} \rangle}{\|\mathbf{g}_v\|_2 \, \|\mathbf{g}_v^{gt}\|_2}, \tag{9.15}$$

(iv) Label loss. When needed, we reuse the SemGaze contrastive objective but apply it only to the winner's label embedding $\mathbf{x}_{k^\star}^{l}$:

$$\mathscr{L}_{\text{lab}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\big(s(\mathbf{x}_{i,k^\star}^{l}, \mathbf{x}_{y_i}^{t})/\tau\big)}{\sum_{c \in \mathscr{C}} \exp\big(s(\mathbf{x}_{i,k^\star}^{l}, \mathbf{x}_{c}^{t})/\tau\big)}, \tag{9.16}$$

where $s(\cdot,\cdot)$ denotes cosine similarity, τ a fixed temperature, and $\mathscr{C}$ the set of unique classes in the batch. Furthermore, $\mathbf{x}_{i,k^\star}^{l} \in \mathbb{R}^{D_l}$ is the predicted gaze label embedding associated with the winning mixture component for sample i.

Model Configuration. The scene image is now processed at 448×448 instead of 256×256, while the head crop is maintained at 224×224. The decoder feature dimension and label embedding size remain $d = 96$ and $D_\ell = 512$, respectively. We use $N_p = 1$ person per image, and set the temperature parameter to $\tau = 0.07$. The *scene encoder* is a DINOv2 ViT-Base, replacing the MultiMAE backbone used in SemGaze. The *gaze encoder* is the same ResNet-18 pretrained

on Gaze360 (Kellnhofer et al., 2019). The probabilistic decoder uses a lightweight MLP that outputs the GMM parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$. A separate MLP with a sigmoid activation predicts the in/out flag.

Training Procedure. We adopt the same optimization setup as SemGaze: AdamW (`lr`=2×10^{-4}, weight decay 0.003), cosine annealing with 4 warmup epochs, total 20 epochs, batch size of 48, and stochastic weight averaging (Izmailov et al., 2018). Loss weights are denoted by λ_{nll}, λ_{dist}, λ_{ang}, and λ_{lab}.

Inference Pipeline. At inference, the label recognition follows the same procedure as SemGaze. We pre-encode a vocabulary of class names with the frozen text encoder and compute similarity scores with the predicted label embeddings to obtain open-vocabulary predictions. For gaze localization, we select the most confident component using a score proportional to the ratio of its mixing coefficient to variance:

$$\hat{k} = \arg\max_{k} \frac{\pi_k}{\sqrt{\det \boldsymbol{\Sigma}_k}}. \tag{9.17}$$

The corresponding mean $\boldsymbol{\mu}_{\hat{k}}$ is reported as the final gaze prediction, and the scalar p_{in} from the in/out head indicates whether the target lies inside or outside the image. This criterion balances mode probability and spatial precision, providing a simple yet effective decoding rule.

9.5 Experimental Design

9.5.1 Benchmark Comparison

Table 9.1 reports results on GazeFollow, VideoAttentionTarget, and ChildPlay, in comparison with prior work up to the time of writing. ProbGaze achieves state-of-the-art performance on all three benchmarks, outperforming existing methods by a clear margin. Notably, unlike competing methods, our model is trained exclusively on GazeFollow and evaluated directly on the video-based datasets without fine-tuning. It is also worth noting that recent approaches such as Yang and Lu (2024) and Ryan et al. (2025) also adopt the same DINOv2 backbone as ours. Interestingly, our model not only surpasses the DINOv2-Base variant used in Ryan et al. (2025), but even outperforms their stronger DINOv2-Large model, despite being trained under a more challenging probabilistic framework based on Gaussian mixtures and maximum-likelihood optimization, rather than direct pixel-wise MSE supervision.

9.5.2 Ablation studies

We conduct a series of controlled ablations to assess the contribution of key design choices in *ProbGaze*. Unless otherwise specified, all experiments use $K = 3$ mixture components,

Method	GazeFollow			VAT		ChildPlay	
	AUC ↑	Avg. Dist. ↓	Min. Dist. ↓	Dist. ↓	AP ↑	Dist. ↓	AP ↑
Recasens et al. (2015)	0.878	0.190	0.113	—	—	—	—
Chong et al. (2018)	0.896	0.187	0.112	0.171	0.712	—	—
Lian et al. (2018)	0.906	0.145	0.081	—	—	—	—
Chong et al. (2020b)	0.921	0.137	0.077	0.134	0.853	—	—
Fang et al. (2021)	0.922	0.124	0.067	0.108	0.896	—	—
Jin et al. (2021)	0.919	0.126	0.076	0.134	0.880	—	—
Jin et al. (2022)	0.909	0.137	0.077	—	—	—	—
Tonini et al. (2022)	0.927	0.141	—	0.129	—	—	—
Bao et al. (2022)	0.928	0.122	—	0.120	0.869	—	—
Hu et al. (2022b)	0.923	0.128	0.069	0.118	0.881	—	—
Gupta et al. (2022)	0.933	0.134	0.071	0.122	0.864	0.113	0.983
GeomGaze (2023)	0.936	0.125	0.064	0.109	0.834	0.107	0.986
Sharingan (2024)	0.944	0.113	0.057	0.107	0.891	0.106	0.990
SemGaze (2024)	—	0.108	0.051	—	—	—	—
Yang and Lu (2024)	0.939	0.114	0.059	0.101	**0.916**	—	—
Song et al. (2024)	0.949	0.105	0.047	0.102	0.905	—	—
Ryan et al. (2025)	0.956	0.104	0.045	0.107	0.897	0.106	0.994
ProbGaze†	—	**0.098**	**0.044**	**0.095**	—	**0.093**	—
Human	0.924	0.096	0.040	0.051	0.925	0.048	0.993

Table 9.1: Results on the GazeFollow, VideoAttentionTarget (VAT), and ChildPlay datasets. We show the best scores in bold. The † symbol means that the model was trained on GazeFollow and evaluated on the video datasets without fine-tuning.

isotropic covariance, and both the NLL and auxiliary distance loss terms. Results are summarized in Tables 9.2–9.6.

Number of Components. Table 9.2 analyzes the impact of the number of Gaussian components K. The single-Gaussian case ($K = 1$) serves as the heteroscedastic regression baseline, which is functionally equivalent to the coordinate regression variant used in *Sharingan* (*cf.* Table 6.2 in Chapter 6), but augmented with an additional variance term. As observed previously, this formulation yields a lower average distance but a slightly higher minimum distance, since the predicted mean tends to collapse the posterior into its expectation.

Increasing the number of components from $K = 1$ to $K = 3$ yields modest improvements in the minimum distance metric, confirming that multiple hypotheses help capture scene-level multimodality when present. However, performance saturates and even slightly declines beyond $K = 3$, likely due to over-parameterization and redundant modes in scenes that rarely contain more than a few plausible gaze targets. Qualitative inspection supports this interpretation: across datasets, $K = 3$ provides a good balance between expressivity and compactness, adequately representing typical one-to-three-target ambiguity without fragmenting probability mass unnecessarily.

Covariance Structure. We further ablate the covariance parameterization, comparing isotropic, diagonal, and full covariance models (Table 9.3). Quantitatively, distance metrics remain

# Modes	Avg. Dist. ↓	Min. Dist. ↓
$K = 1$	0.083	0.047
$K = 2$	0.097	0.046
$K = 3$	0.098	0.044
$K = 5$	0.101	0.047

Table 9.2: Number of components ablation.

Covariance	Avg. Dist. ↓	Min. Dist. ↓
Full	0.098	0.044
Diagonal	0.098	0.045
Isotropic	0.098	0.044

Table 9.3: Covariance matrix ablation.

Losses	Avg. Dist. ↓	Min. Dist. ↓
$\mathscr{L}_{\text{nll}}$	0.097	0.046
$\mathscr{L}_{\text{nll}} + \mathscr{L}_{\text{dist}}$	0.098	0.044

Table 9.4: Losses composition ablation.

Decoder	Avg. Dist. ↓	Min. Dist. ↓
Heatmap	0.094	0.040
GMM	0.098	0.044

Table 9.5: Decoder type ablation.

3D FoV	Avg. Dist. ↓	Min. Dist. ↓
w/o FoV	0.102	0.048
w/ Fov	0.098	0.044

Table 9.6: Field-of-view ablation.

Rule	Avg. Dist. ↓	Min. Dist. ↓
ML(x)	0.096	0.050
$\pi \uparrow$	0.098	0.046
$\sigma \downarrow$	0.098	0.045
$\frac{\pi}{\sigma^2} \uparrow$	0.098	0.044

Table 9.7: Decision rule ablation.

largely unchanged, indicating that the added expressivity of richer covariance structures offers little benefit while increasing computational and numerical complexity. However, distance-based metrics may not fully capture the value of uncertainty modeling, as they evaluate only point accuracy rather than distributional quality. Overall, the isotropic formulation provides an effective and efficient approximation for gaze prediction, balancing simplicity and robustness for downstream use.

Loss Composition. Table 9.4 analyzes the effect of adding the winner-takes-all (WTA) distance loss on top of the negative log-likelihood. While $\mathscr{L}_{\text{NLL}}$ alone optimizes the likelihood of the ground-truth point, it tends to under-emphasize precise localization for uncertain samples due to the variance term in the denominator. Adding $\mathscr{L}_{\text{dist}}$ counteracts this by explicitly enforcing spatial accuracy on the best-explaining component, mirroring the rebalancing strategies proposed in prior work on faithful heteroscedastic regression (Seitzer et al., 2022; Stirn et al., 2023). The combined objective yields slightly sharper predictions and a lower minimum distance error.

Decoder Type. We compare the proposed probabilistic decoder against a traditional heatmap-based formulation (Table 9.5). The heatmap baseline corresponds to the same architecture introduced in *SemGaze* (Chapter 8), but re-evaluated here with a stronger visual backbone (DINOv2), higher input resolution (448×448), and the addition of 3D field-of-view features. Despite relying on a much simpler prediction head (i.e. an MLP that outputs mixture parameters instead of a spatial upsampling decoder), the GMM variant performs nearly on par

with the heatmap model, while being more compact and offering interpretable uncertainty estimates.

Interestingly, the heatmap model itself appears to have reached the performance ceiling of the GazeFollow benchmark, achieving results that closely match reported human-level accuracy (0.094 vs. 0.096 in average distance and 0.040 vs. 0.040 in minimum distance). This underscores the challenge of further improving purely spatial formulations and highlights the value of probabilistic representations that capture uncertainty rather than solely minimizing point-wise error.

Effect of 3D Field-of-View. Finally, Table 9.6 quantifies the benefit of incorporating the 3D FoV representation, which integrates scene depth into the gaze reasoning process. Adding this geometric cue consistently improves both average and minimum distance metrics, confirming that depth-aware representations help rule out gaze candidates that do not match the 3D field of view of the subject. The same improvement is also found in ChildPlay (distance of 0.093 vs. 0.102) and VideoAttentionTarget (distance of 0.095 vs. 0.098). This phenomenon was already highlighted in Chapter 5.

Decision Rules. Table 9.7 compares different strategies for selecting the final gaze point from the Gaussian mixture output. All rules yield similar performance overall, with small but consistent improvement for the decision rule based on the most likely mode (i.e. $\arg\max_k \frac{\pi_k}{\sigma_k^2}$). The maximum-likelihood point (ML(x)) achieves the lowest avg_dist (0.096) but a noticeably higher min_dist (0.050), indicating that it tends to predict more central, averaged locations. This point corresponds to the global mode of the mixture, obtained via the mean-shift fixed-point iteration:

$$x^{(t+1)} = \sum_k \tilde{r}_k(x^{(t)})\,\mu_k, \quad \tilde{r}_k(x) = \frac{\frac{\pi_k}{\sigma_k^2}\,\mathcal{N}(x \mid \mu_k, \sigma_k^2 I)}{\sum_j \frac{\pi_j}{\sigma_j^2}\,\mathcal{N}(x \mid \mu_j, \sigma_j^2 I)}.$$

This update acts as a *soft averaging mechanism* over nearby components, naturally pulling the solution toward the center of overlapping Gaussians and thus aligning better with the average gaze annotation. Intuitively, the mixture mode seeks to *explain all nearby hypotheses jointly* (i.e. a compromise that balances the influence of multiple clustered gaze targets), whereas the "most-likely component" ($\frac{\pi}{\sigma^2}\uparrow$) selects one hypothesis decisively. Therefore, even though we refer to ML(x) as the "maximum-likelihood" point, it is not guaranteed to coincide with any individual mode: when components overlap, it can fall between peaks where the combined mixture density is higher.

9.5.3 Uncertainty Calibration

To assess whether the predicted uncertainties are meaningful, we analyzed their correlation with empirical gaze errors. Figure 9.5 plots the average minimum distance to the ground-truth annotations against two complementary uncertainty measures derived from the predicted

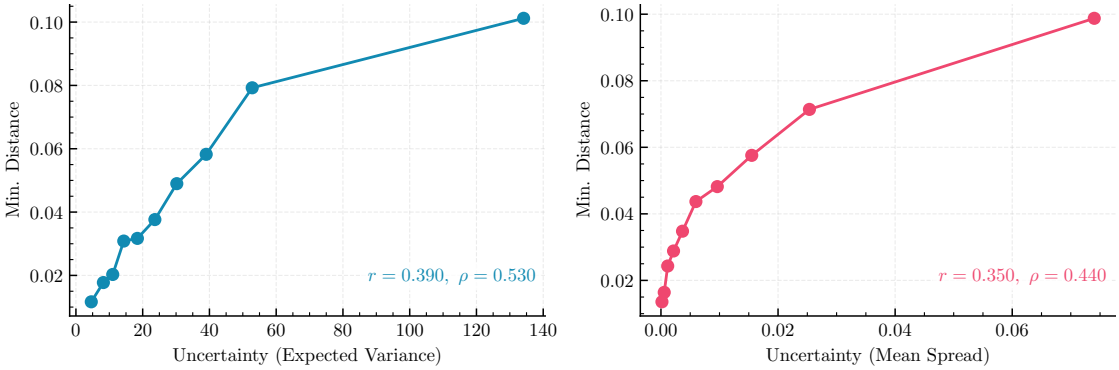


Figure 9.5: Calibration of predicted uncertainties. **Left:** Within-mode uncertainty, measured as the mixture-weighted average variance $\sum_{k=1}^{K} \pi_k \operatorname{tr}(\boldsymbol{\Sigma}_k)$. **Right:** Between-mode uncertainty, measured as the spatial spread of component means $\sum_{k=1}^{K} \pi_k \|\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}\|^2$ where $\bar{\boldsymbol{\mu}} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$. For reference, we also include Pearson's correlation r and Spearman's ρ. Both measures show a conclusive monotonic relationship with empirical gaze error, indicating that the predicted uncertainties of *ProbGaze* are well calibrated: samples with larger predicted uncertainty (in the form of variance or mode spread) systematically correspond to larger prediction errors. Note that $\boldsymbol{\Sigma}_k$ is expressed in pixel space (64×64 heatmap resolution), while $\boldsymbol{\mu}_k$ is defined in normalized image coordinates ($[0,1] \times [0,1]$), hence the two quantities differ in numerical scale but remain comparable in trend.

Gaussian Mixture Model (GMM). The left panel reports the expected within-mode variance, which quantifies local uncertainty around each predicted gaze mode, while the right panel shows the between-mode variance (mean spread), which captures global ambiguity across multiple gaze hypotheses. In both cases, the curves exhibit a clear, monotonic increase, where samples with larger predicted variance correspond systematically to larger gaze errors (Pearson $r = 0.39$, Spearman $\rho = 0.53$ for within-mode variance; $r = 0.35$, $\rho = 0.44$ for mean spread). This consistency indicates that the model's predicted uncertainties are well-calibrated: they provide a reliable proxy for the true prediction difficulty. Together, the two measures reveal that both local dispersion and multimodal ambiguity contribute meaningfully to uncertainty in gaze target prediction.

9.5.4 Multi-Hypothesis Semantic Embeddings

Table 9.8 compares the performance of *SemGaze*, *ProbGaze (SE)*, and the full *ProbGaze* model on the GazeFollow dataset. The *ProbGaze (SE)* variant serves as a control version that uses the same single semantic embedding as *SemGaze*, allowing us to isolate the effects of architectural and training differences introduced in this chapter (i.e. the stronger DINOv2 backbone, higher input resolution (448×448), and the integration of 3D FoV features).

In terms of localization, both *ProbGaze* and its single-embedding counterpart achieve similar accuracy, clearly outperforming *SemGaze* thanks to the improved backbone, resolution, and geometric conditioning. For recognition, however, the differences are more revealing. Switching from *SemGaze* to *ProbGaze (SE)* yields a notable improvement of +5.2 Acc@1 and +5.6 SetAcc@1 points, driven by the architectural refinements above and by fixing the temperature parameter in the contrastive loss, which we found to significantly stabilize training and

Method	Localization		Recognition		
	Avg. Dist. ↓	Min. Dist. ↓	Acc@1 ↑	Acc@3 ↑	SetAcc@1 ↑
Random	0.471	0.391	0.002	0.010	0.003
Bias / Majority	0.295	0.229	0.010	0.015	0.011
SemGaze (2024)	0.108	0.051	0.393	0.581	0.452
ProbGaze (SE)	0.099	0.045	0.445	0.650	0.508
ProbGaze	**0.098**	**0.045**	**0.493**	**0.697**	**0.571**

Table 9.8: Results for gaze target localization and semantic recognition on the GazeFollow dataset. SE stands for single-embedding for the semantic gaze label.

improve alignment quality.

A key limitation of single-embedding formulations is that they conflate uncertainty from two distinct sources: (*i*) uncertainty over the gaze target's location (e.g. which object the person is looking at) and (*ii*) uncertainty over the target's semantic class (e.g. what that object is). When the gaze is ambiguous, the model averages over spatially distinct possibilities, and this "mode averaging" propagates into the semantic embedding, blurring class distinctions. The proposed multi-hypothesis formulation mitigates this by assigning a separate embedding to each Gaussian component, each conditioned on the spatial mean and variance of that mode. This design effectively suppresses the uncertainty arising from localization while preserving the semantic uncertainty, allowing cleaner, more decisive class representations. Empirically, this leads to a marked gain of +4.8 Acc@1 and +6.3 SetAcc@1 improvement compared to the single-embedding variant.

Due to hardware constraints, the semantic contrastive loss was trained with a batch size of 48 for all three models (versus 300 in the original *SemGaze* configuration), limiting negative diversity and slightly underestimating potential performance. Nevertheless, our quantitative analysis demonstrates the benefit of multi-hypothesis prediction, and the qualitative examples in Figure 9.6 further illustrate this effect: each Gaussian mode corresponds to a distinct spatial location with a semantically meaningful embedding corresponding to its underlying object.

9.5.5 Qualitative Evaluation

Figure 9.7 presents representative predictions of *ProbGaze* across GazeFollow, VideoAttentionTarget, and ChildPlay, along with typical failure cases. The second column shows the predicted *3D field of view* (FoV) for a selected subject, while the third column visualizes the predicted *GMM* as a heatmap.

The predicted mixture densities are remarkably well calibrated. In unambiguous scenes (rows 2–4), multiple components converge toward a single dominant mode, while in uncertain scenes (rows 1, 5, and 6), the distribution spreads over multiple plausible gaze targets. Beyond the mean locations, the predicted variances also carry semantic meaning: for instance, on large or textureless surfaces, the model produces higher variance estimates, reflecting genuine

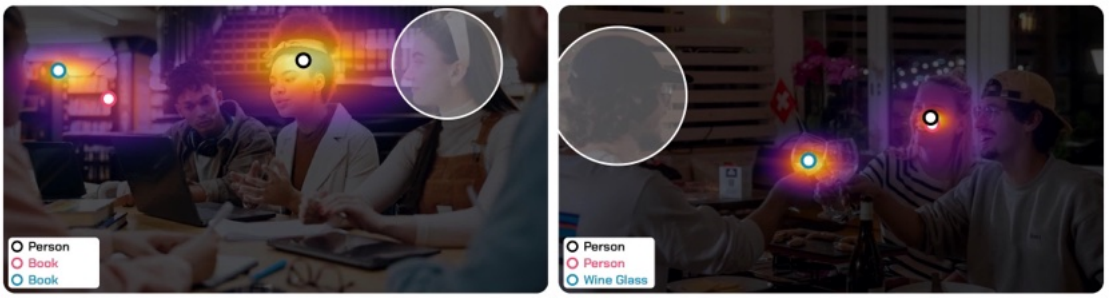


Figure 9.6: Qualitative examples of multi-modal gaze predictions from *ProbGaze* on internet images. Each spatial mode predicted by our model is associated with its own semantic embedding (and class), enabling the model to describe the underlying object corresponding to that mode. Even when the scene admits multiple semantically distinct gaze targets (e.g. different objects), the model assigns a separate embedding to each spatial hypothesis, linking spatial and semantic uncertainty in a unified representation. The Gaussian mixture is visualized as a heatmap, and the mode selected as the final gaze prediction is indicated by a black border. We employ the vocabulary of the COCO dataset (80 object categories).

uncertainty about where exactly the gaze may land (rows 1 and 6). The 3D FoV representation further proves effective at suppressing implausible targets that lie at inconsistent depth levels (row 4), confirming the geometric consistency of the approach.

Failure cases generally fall into two categories. The first involves mild localization errors, where the predicted mean is slightly off target (row 7). The second involves selecting the wrong gaze hypothesis entirely (row 8), typically due to an inaccurate 3D gaze vector estimation that misguides the FoV toward irrelevant regions.

Finally, these visualizations suggest that simple post-processing could further refine predictions. When the scene is unambiguous, the model naturally collapses multiple components onto a single position or assigns very low weights to redundant modes. This structure could be exploited by pruning low-confidence components and merging nearby ones before re-normalizing the mixture coefficients. This should yield a more compact, interpretable posterior without retraining.

9.6 Limitations

While the proposed probabilistic formulation provides a compact and principled way to model gaze uncertainty, several limitations remain.

Single-point Supervision. Although the maximum likelihood objective naturally supports multiple annotations per sample, current datasets provide only one annotation per person, which limits how well multimodality can be learned. Collecting multiple gaze annotations per sample (concentrated on a single target in clear cases and distributed across plausible ones in ambiguous scenes), would yield a more faithful training signal.

Semantic Supervision. The semantic supervision available in GazeFollow is limited in both

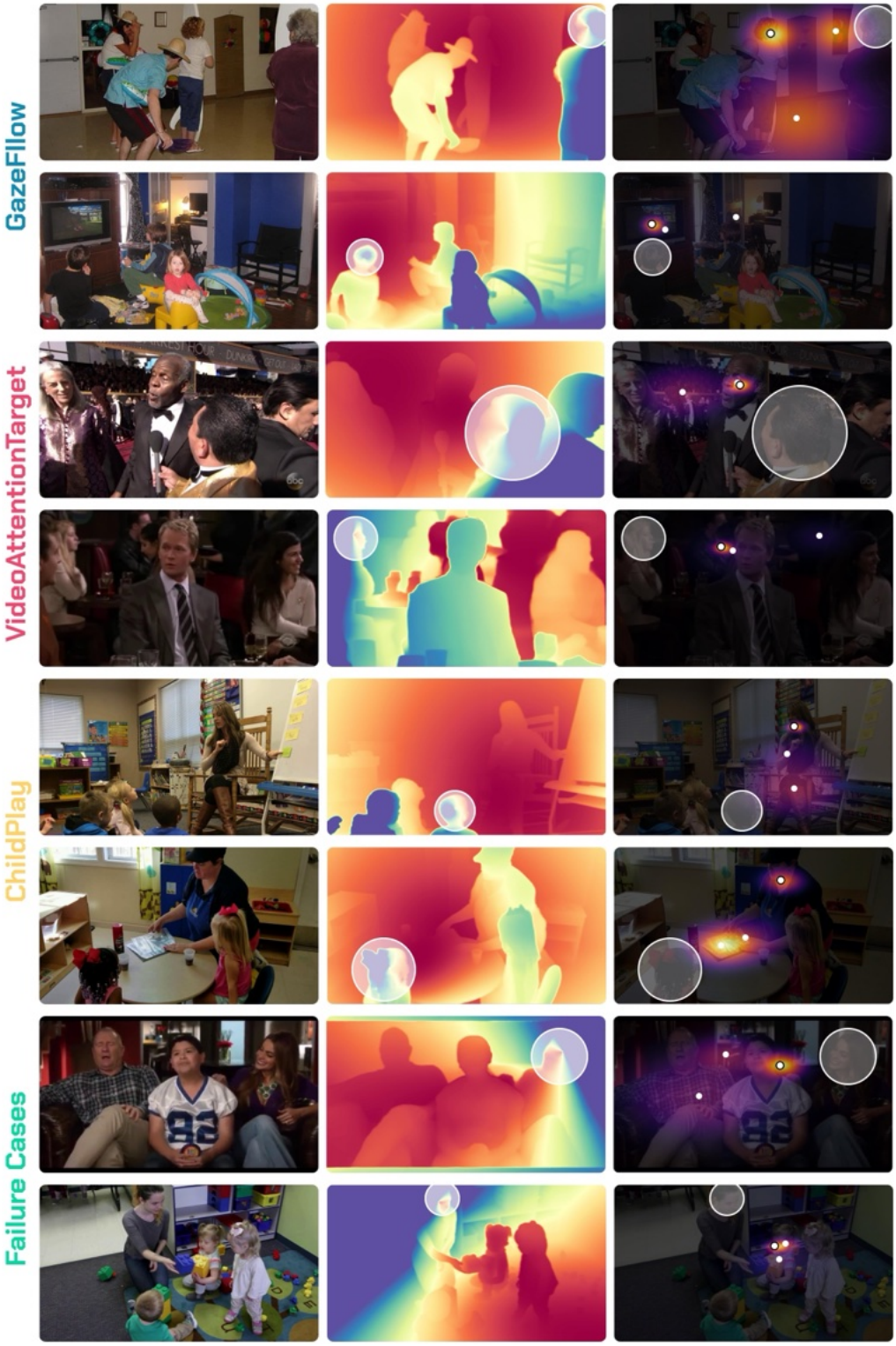


Figure 9.7: Predictions of ProbGaze on images from standard benchmarks. Each row shows the input image, the corresponding 3D field-of-view (FoV) heatmap for a selected person, and the predicted Gaussian mixture distribution visualized as a gaze heatmap. Mixture component means $\boldsymbol{\mu}_k$ are depicted as white circles, with the selected gaze mode outlined in black. We use a model trained on GazeFollow.

scale and diversity. Its 100K pseudo-labeled images are insufficient to train a robust and knowledgeable open-world semantic tagger. Moreover, each image typically contains several identifiable objects, yet supervision is provided only for the one corresponding to the annotated gaze target. Since the predicted semantic embeddings in our model depend solely on the frozen scene tokens and the spatial parameters ($\boldsymbol{\mu}, \boldsymbol{\sigma}$), an interesting direction for future work is to augment training by leveraging other objects present in the scene. Specifically, one could sample points within their segmentation masks as pseudo gaze targets (using the mask centroid as $\boldsymbol{\mu}$ and a standard deviation covering their extent), thereby aligning the semantic embeddings with a richer set of scene entities beyond the ground-truth gaze target.

Output Representation Gap. A small but consistent performance gap remains between the GMM and the heatmap decoder. While the mixture formulation is far more compact and uncertainty-aware, further work is needed to close this gap, possibly through improved loss formulations, calibration techniques, or regularization schemes inspired by the broader mixture density literature.

Depth Dependency. The integration of the 3D Field-of-View requires a depth map at training and inference time, which adds computational overhead and dependency on depth estimation quality. One potential direction is to adopt a "FoV dropout" strategy, training the model to operate both with and without depth cues, thereby increasing robustness to missing or unreliable geometry. The FiLM modulation scheme used in our method naturally supports this gating mechanism, allowing the network to down-weight or ignore the FoV features when they are noisy or unavailable.

9.7 Conclusion

In this chapter, we reframed gaze prediction not as a coordinate regression or dense heatmap estimation problem, but as the modeling of a probability distribution over possible gaze targets. The contribution lies less in architectural novelty or performance gains, and more in the representational shift toward a compact, interpretable, and flexible probabilistic formulation. By predicting full distributions rather than deterministic outputs, the model learns to express uncertainty in a way that mirrors human reasoning under ambiguity: being decisive when the scene is clear and probabilistic when multiple hypotheses coexist.

This probabilistic treatment combined with the promptable architecture design provides a unifying bridge between geometry and semantics, integrating spatial reasoning with contextual interpretation within a single, likelihood-based framework. It also lays the foundation for future exploration of uncertainty-aware perception systems that reason not only about what they see, but also about how confident they are in what they infer.

The next part, *Discussion and Outlook*, moves beyond methods to examine broader implications. It addresses the ethical, practical, and conceptual dimensions of gaze modeling, and charts open challenges and future directions for gaze understanding in-the-wild.

Part III Discussion & Outlook

10 Limitations & Ethical Considerations

Despite steady advances in architecture design and benchmarking, gaze following systems still face structural limitations that constrain both scientific insight and real-world impact. This chapter surveys those limitations along three methodological fronts: data, models, and evaluation, highlighting how small, narrowly curated datasets and shallow annotations bias learning; how current models over-rely on geometric shortcuts while under-utilizing social and semantic cues; and how fragmented protocols obscure genuine progress. Finally, we discuss ethical and privacy considerations intrinsic to collecting and sharing personally identifiable video data, and close with broader reflections on when gaze following is genuinely useful, advocating that utility should guide future work, instead of pixel-perfect determinism.

10.1 Methodological Limitations

Despite the steady progress of gaze-following research over the past decade, fundamental methodological limitations persist. These limitations can be grouped into three categories: data, model design, and evaluation protocols.

10.1.1 Data-related Limitations

The ability of gaze-prediction models to generalize and reason about attention is ultimately constrained by the nature of the data they are trained on. Most publicly available datasets are small in scale, narrow in context, and shallow in annotation.

Limited scale, diversity, and annotation quality. Even the largest video-based benchmarks (e.g., VideoAttentionTarget, ChildPlay) comprise on the order of only 100–300 distinct scenes, far from sufficient to capture the diversity required to model human gaze behavior. Even large-scale image datasets such as GazeFollow exhibit issues of noisy annotations, biased subject sampling, and narrow spatial context, leaving a considerable gap from the richness and unpredictability of *truly in-the-wild* scenes.

Missing layers of annotation. Current datasets focus mainly on where people look (2D points) but seldom encode what or whom they look at, why, or under what social configuration. Important complementary layers of annotation are often absent:

- *Object-level* information (bounding boxes or instance masks) to link gaze points to physical entities.
- *Semantic classes* describing countable objects and uncountable "stuff" regions (e.g., wall, floor, sky).
- *Social labels*, such as speaker/listener roles, engagement links, or activity categories that would allow conditioning gaze reasoning on interaction context.

Without these complementary cues, models must infer them implicitly from geometry and appearance alone, which leads to ambiguous or brittle reasoning.

10.1.2 Model-related Limitations

While humans integrate a wide spectrum of cues to infer gaze (e.g. geometry, motion, social roles, speech, and activity), current models rely mostly on geometric direction and saliency.

Shortcut learning. My experience across multiple architectures suggests that models achieve apparently good results by exploiting superficial correlations between head orientation and salient regions. When several plausible gaze targets exist, a human observer uses subtle cues (e.g. lip movement, gesturing, conversational turn-taking), to disambiguate. Current deep models are blind to these cues and default to geometric and saliency shortcuts. They "solve" the benchmark but not the underlying cognitive task. This tendency of deep networks to rely on features that are *accessible* rather than *causally predictive* is well documented in the literature (e.g. Hermann et al. (2023)).

The cost of implicit learning. Capturing such rich contextual understanding implicitly would require astronomical data scale and exposure to an immense variety of real-world interactions. Instead, a more efficient path forward is to explicitly annotate key social or behavioral signals and use them to regularize training:

- Auxiliary supervision for roles (speaker, listener), engagement, or activity.
- Auxiliary tasks like predicting who interacts with whom or what action is being performed. This supervision gives the model an anchor to interpret geometry within a social context, rather than rediscovering these concepts unsupervised.

In fact, many of these ideas are already recognized within the research community and have been explored to varying degrees. For instance, the work of Hou et al. (2024) explicitly leverages speech cues, noting that being the main speaker in a group markedly increases the likelihood of being the gaze target of others (Ba and Odobez, 2010; Siegfried and Odobez, 2021). A more comprehensive treatment of gaze behavior is found in Wei et al. (2018), which considers not only human attention but also underlying intentions and ongoing tasks. However, their

dataset is scripted and has not been publicly released, limiting its practical impact. Overall, such efforts remain fragmented and lack unification: each focusing on isolated sub-tasks rather than integrating them within a consistent framework or common evaluation protocol.

Self-supervised and masking objectives. A complementary strategy is to enforce contextual reasoning by design. For example, masking the head or eyes and training the model to infer the gaze target compels it to rely on surrounding cues (body posture, social layout) rather than the head itself. Such self-supervised objectives can act as a proxy for learning human-like context sensitivity without explicit labels.

Inter-person interactions. Even in transformer-based multi-person architectures, inter-person communication remains weakly modeled. Humans naturally interpret gaze relationally: each individual's attention constrains the likely targets of others. Encoding such relational dependencies may require graph-based or message-passing mechanisms coupled with semantic grounding (e.g. recognizing who is observing whom or which group is sharing attention).

One experiment we conducted while investigating the *Sharingan* architecture from Chapter 5 was to analyze the attention weights between gaze tokens as a proxy for model-based interaction. We observed that gaze tokens tend to attend to one another when the underlying people are engaged in a shared attention behavior. However, this should not be interpreted as evidence of genuine social understanding. The effect is largely driven by geometric alignment: when multiple people look at the same location, their corresponding tokens must ultimately encode similar spatial information, and thus naturally attend to each other. This also explains why the same attentional pattern was not observed in cases of mutual gaze, where people look at each other rather than a shared point (i.e. the underlying spatial targets differ, and so does the induced token interaction).

10.1.3 Evaluation-related Limitations

Even when datasets are adequate, evaluation practices remain inconsistent and incomplete, limiting our ability to measure genuine progress.

Inadequate metrics. Most benchmarks rely on pixel-level metrics such as 2D distance or AUC, which ignore the semantic correctness of the prediction. A gaze point that falls one pixel inside a table versus one pixel on a person's face may have vastly different meanings, yet both count equally. Without object-level or semantic annotations, these metrics provide little insight into what a model has understood. Metrics such as Semantic Gaze Accuracy (SGA, used in Chapter 8) are a first step toward solving this problem, but datasets with object-level annotations remain scarce.

Fragmented evaluation protocols. A persistent challenge in gaze-following research is the lack of standardized evaluation protocols. Many papers introduce their own training–testing splits

or variant procedures, even when using the same benchmarks, making cross-comparison both difficult and unreliable. This fragmentation obscures real progress and encourages models to overfit to specific datasets rather than promoting genuine generalization. The issue is particularly acute in social gaze tasks, where sub-tasks such as shared attention and mutual gaze are often treated in isolation instead of being evaluated within a unified framework that considers accuracy, semantics, and social context jointly.

A striking example is the one-stage formulation of gaze target prediction proposed by Tu et al. (2022), where the model jointly predicts head bounding boxes and corresponding gaze targets. The difficulty lies in the fact that existing benchmarks do not annotate all people present in the scene, which makes it impossible to properly compute precision and recall. As a result, the evaluation protocol introduced in Tu et al. (2022), and later adopted in Tonini et al. (2023), *artificially inflated performance metrics*, producing scores so high that no standard two-stage approach could appear competitive by comparison.

This situation is problematic because it risks distorting progress and misguiding the community. In our own work, we spent a significant amount of time clarifying to reviewers why two-stage methods should not be compared under that protocol. Moreover, newcomers to the field may fail to recognize these discrepancies and misinterpret the results, as seen in Song et al. (2024). Consequently, several recent papers, including our own (Tafasca et al., 2024b; Gupta et al., 2024b; Ryan et al., 2025), have begun explicitly addressing this issue in their methodology sections to preempt confusion and encourage fair, transparent evaluation practices.

Toward holistic evaluation. More fundamentally, gaze understanding should move toward unified, multi-task evaluation. In real applications (e.g. robotics, social analysis, or developmental screening), it is rarely sufficient to predict only the gaze point or only the eye-contact state. These cues co-occur and jointly define meaning. Evaluating models across isolated sub-tasks (e.g. gaze following, LAEO, shared attention) overlooks this interdependence. For example, in formal autism diagnoses, the ADOS assessment protocol explicitly defines the *Initiation of Joint Attention (IJA)* event as a 2-step or 3-step process where the patient alternates between making eye-contact with the clinician and looking at a toy (*cf.* Figure 10.1).

A unified framework should therefore:

- Benchmark models across gaze localization, social gaze, and semantic interpretation simultaneously.
- Encourage multi-task or multi-behavior evaluation, measuring how well models maintain consistency across these tasks.
- Be grounded in large-scale, unified datasets where all relevant labels (i.e. gaze point, target class, social relations) coexist within the same scenes.

Such a holistic approach would move the field from narrow metric optimization toward genuine understanding of visual attention in everyday scenes.

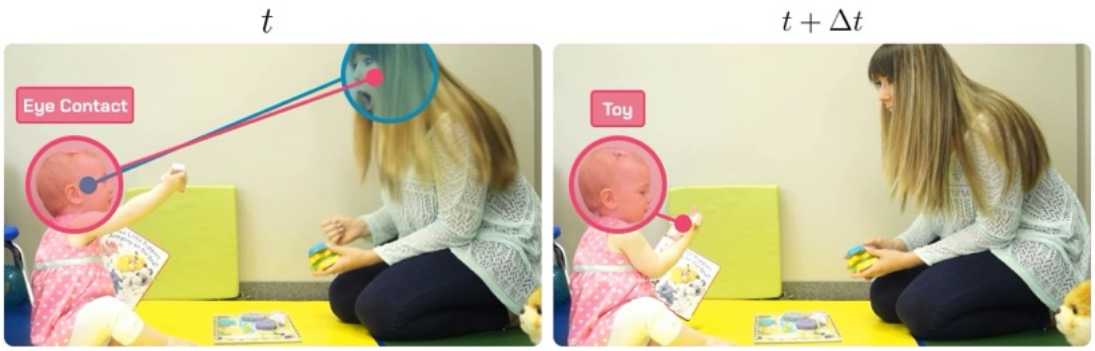


Figure 10.1: Illustration of a two-step initiation of joint attention. The child first establishes eye contact with the woman to invite her to share attention toward the toy, then looks at the toy in her hand. Recognizing such gaze events requires a model capable of localizing and semantically identifying gaze targets, understanding social interactions, and reasoning about these cues over time.

10.2 Ethical & Privacy Considerations

Modern gaze-following research operates at the intersection of computer vision and human behavior understanding. As the field advances toward ever more realistic and socially aware systems, ethical reflection and privacy protection become integral components of responsible development.

10.2.1 Handling Personally Identifiable Visual Data

Most gaze datasets and models rely on raw RGB imagery in which faces, eyes, and even children's identities are clearly visible. Such data are by nature personally identifiable and often fall under strict data-protection regulations (GDPR, HIPAA, etc.). The situation is particularly sensitive when recordings involve minors, clinical contexts, or domestic environments. Collecting, storing, and sharing such data therefore requires explicit informed consent, secure storage protocols, and preferably on-site processing to prevent raw data from leaving institutional servers. In this light, an emerging research direction focuses on feature-based gaze prediction—learning from modalities such as pose, depth, or semantics instead of facial appearance. Our own work on multimodal gaze following (Gupta et al., 2022) demonstrated that combining these modalities retains acceptable gaze prediction accuracy while discarding identifiable image regions. More recently, Toaiari et al. (2024) expanded this idea to 3D settings, replacing head crops with upper-body skeletons and depth maps to preserve identity while still achieving state-of-the-art performance. These approaches illustrate that privacy preservation and performance need not be mutually exclusive.

10.2.2 Risk *vs.* Utility

Like many perceptual technologies, gaze analysis sits at a moral crossroads. On one hand, it offers substantial social value: aiding autism research, enabling assistive robotics for people with disabilities, and supporting clinical assessment or educational tools that rely on

understanding attention. On the other hand, the very same capability can be repurposed for surveillance, targeted advertising, or attention profiling, which are all applications that intrude on autonomy and raise legitimate societal concern.

The contrast is particularly stark when comparing therapeutic and commercial settings. In autism therapy, gaze analysis provides clinicians with non-invasive behavioral insight that would otherwise require hours of manual coding. In retail analytics, however, using gaze to quantify consumer interest in products or store layouts risks reducing human attention to a commodified metric—the caricature of late-stage capitalism. The technology itself is *neutral*; its moral valence depends entirely on the context, intent, and governance mechanisms surrounding its deployment.

10.2.3 Toward Privacy-Preserving Design Principles

Balancing utility and risk demands privacy by design. Three complementary strategies are emerging:

- Modality substitution: Replace identifiable RGB input with abstract features (pose, depth, segmentation, or semantic maps), or use generative anonymization techniques (e.g. replacing real faces with synthetic but geometrically consistent surrogates generated by diffusion or GAN-based models). These methods preserve the structural information required for gaze inference while removing personally identifiable facial traits. However, they must be applied transparently and ethically, as generative anonymization itself can introduce artifacts or biases if not carefully controlled.
- Data minimization: Process data locally and retain only aggregated or anonymized gaze descriptors.
- Ethical transparency: Clearly state intended use, potential misuse, and limitations in all dataset and model releases.

By shifting representation from who the person is to what they are doing, gaze understanding systems can move away from encoding identity-specific features (e.g. facial appearance, hairstyle, or clothing) toward representations that emphasize behavior, context, and intent. In practice, this means replacing or transforming appearance-rich inputs with behavioral abstractions. For example, representing people through their pose skeletons, depth silhouettes, or semantic segmentation masks instead of raw face images (*cf.* Figure 10.2). At a deeper level, this shift can also be enforced through *representation-level disentanglement*, where the model learns to separate identity-related information from task-relevant behavioral cues. Adversarial or contrastive objectives can, for instance, explicitly suppress identity signals in latent features while preserving those related to gaze direction, motion, or social engagement (e.g. Chen et al. (2018)). In doing so, models learn to reason about how a person behaves, looks, or interacts, rather than focus on their identity. This not only mitigates privacy risks but also promotes generalization, since behavior tends to transfer across individuals more consistently than appearance. However, adversarial or contrastive disentanglement does not

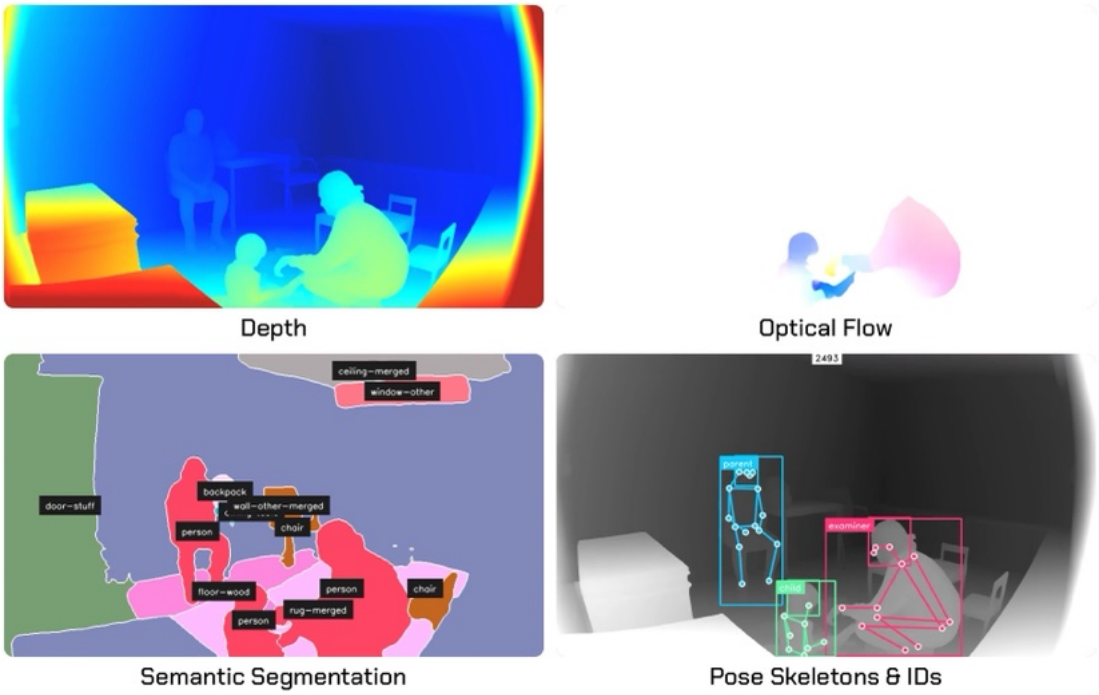


Figure 10.2: Privacy-preserving multimodal representation of an autism diagnosis session (ADOS). Instead of distributing raw RGB footage, the session is represented through multiple derived modalities that capture complementary behavioral and contextual information: depth maps (geometry), optical flow (motion), semantic segmentation (scene and object understanding), pose skeletons (body configuration and social layout), and tracking (ID consistency). Additional features such as image embeddings, gaze or gesture estimates, and other behavioral descriptors can further enrich these representations. This approach enables the sharing of sensitive clinical data with the research community while preserving privacy, facilitating progress in areas where data access has long been a major bottleneck.

make the *training process* itself private; rather, it makes what the model learns and outputs more privacy-preserving. In other words, these methods mitigate leakage from the model's representational space, acknowledging that the model itself lies within the privacy threat surface.

In the long term, we may also envision gaze understanding systems that adaptively choose privacy levels: using full-image cues only when necessary and otherwise relying on less intrusive modalities.

10.3 Broader Limitations of Gaze Following as a Task

Despite its appeal as a window into human attention and interaction, gaze following remains an inherently **ill-posed problem**. The visual signal alone is often ambiguous, and even humans cannot always disambiguate where someone is looking. This limitation gives rise to what can be described as *the illusion of determinacy*: the comforting but misleading assumption that every gaze direction has a single correct and observable target. In reality, gaze is frequently diffuse, shared among several possible referents, or directed toward regions outside the field



Figure 10.3: Contexts where gaze prediction is useful versus non-informative. **Left:** a close, interaction-centered scene where people, objects, and social cues are clearly visible, allowing gaze models to make meaningful inferences about attention and intent. **Right:** a wide surveillance view in a shopping mall, where people appear distant and the visual evidence is insufficient to reliably infer where they are looking. In such unconstrained settings, gaze prediction becomes possible but largely uninformative.

of view. For a learning algorithm, attempting to model such inherently uncertain cases risks turning the problem into an exercise in overfitting to ambiguity.

A more constructive view recognizes that **not all gaze events are equally informative**. Some moments carry clear communicative or task-related intent (e.g. a child following an examiner's gesture or a robot anticipating a human partner's next action), while others are random, idle, or ambiguous. This distinction echoes the broader principle that models are not meant to be correct, they are meant to be useful. As George Box famously put it, *"All models are wrong, but some are useful"*. In this light, the goal of gaze modeling should not be to eliminate uncertainty, but to identify and exploit informative instances where gaze serves as a reliable proxy for cognition or intention. This connects directly to the discussion of predictive uncertainty in Chapter 9, where uncertainty quantification can be used to help discriminate between meaningful and non-meaningful predictions. By acknowledging that some forms of uncertainty are intrinsic (aleatoric) and unavoidable, we can redirect our efforts toward capturing the moments that matter most.

The practical implication is that gaze following is only as valuable as the context in which it is applied. While the formulation promises generalized gaze understanding from exocentric scenes, the visual evidence in many real-world situations is insufficient to resolve precise gaze targets, limiting the utility of such predictions. The question, then, is not whether gaze following can be done everywhere, but *when and where* it is meaningful to do so. In close, interaction-centered scenarios (e.g. therapy sessions, collaborative robotics, or small-group conversations), the visual constraints and social cues make gaze modeling both feasible and informative. In contrast, in distant or large-scale surveillance settings (e.g. overhead cameras in shopping malls), the task loses interpretability: even humans fail to infer gaze direction or target in these scenes.

A useful way to visualize this distinction is to compare two examples: a close-up interaction, where people's heads, hands, and targets are clearly visible and their intentions discernible,

versus a wide surveillance shot where individuals appear small and indistinct (*cf.* Figure 10.3). The former represents the domain where gaze following is useful; the latter, where it is possible but meaningless. This observation also helps explain why most research and funding in gaze understanding come from robotics and autism research, fields where scenes are structured, controlled, and socially rich.

Ultimately, the true value of gaze prediction is not in achieving pixel-perfect accuracy but in its correlation with cognition, communication, and interaction. By treating gaze as an informative behavioral signal rather than a deterministic label, we move closer to models that are not only technically sound but also socially and cognitively relevant.

11 Open Challenges and Future Directions

11.1 Building Richer Datasets

Progress in gaze prediction has long been constrained by the limitations of existing datasets. As discussed in the previous chapters, most benchmarks are small, lack diversity, and provide shallow supervision: typically a 2D gaze point and an in/out label. Moving forward, the field should invest in building datasets that are richer in annotation layers, broader in coverage, and smarter in collection strategy, similar in spirit to the VSGaze dataset we introduced in Gupta et al. (2024b).

Multi-layer and semantically aware annotations. Future datasets should move beyond single-layer annotations to capture the multiple cues involved in gaze reasoning. In addition to gaze point and head box, datasets could include object or person boxes or masks (e.g. GazeHOI from Chapter 3), 3D geometry, role labels (e.g. speaker, listener), and interaction states (joint attention, manipulation). Such multi-layer supervision enables learning models that not only localize gaze but also understand why attention is directed to a specific target.

Synthetic and simulated data. Another promising direction is the controlled generation of gaze data through simulation or generative approaches. Recent work in human-object interaction like SynHOI (Yang et al., 2023) demonstrates how physically plausible, photo-realistic synthetic data can complement scarce real annotations. Synthetic pipelines could allow sampling from a wider range of viewpoints, object layouts, and social configurations that are rarely observed in natural videos. These datasets can expose models to broader data coverage and higher annotation quality, effectively improving performance and generalization.

Active learning and human-in-the-loop annotation. Because dense gaze labeling is expensive, efficient annotation will be crucial for scaling. Active-learning approaches (Li et al., 2024) can prioritize the most informative samples by estimating epistemic uncertainty rather than random sampling. However, this remains challenging in gaze prediction, where much of the uncertainty is aleatoric, arising from inherent ambiguity. A model must therefore sep-

arate what it does not know (epistemic) from what cannot be known (aleatoric) to identify which samples merit annotation. The work of Tonini et al. (2024) explores an active learning approach to reduce the burden of data annotation in gaze following. They propose a novel acquisition function based on three different metrics combining intermediate model activations, output heatmaps, and scene content. Beyond active learning, human-in-the-loop annotation frameworks (Wu et al., 2022) could further help by combining automatic scoring with selective expert review, enabling iterative dataset refinement and reducing annotation cost.

Evaluation and generalization. Richer datasets should also encourage more holistic evaluation. Traditional metrics like distance remain necessary but insufficient; object-centric metrics (e.g. Gaze Object Accuracy) or social metrics provide complementary insights into how well models capture social and communicative aspects of gaze. Furthermore, dataset design should explicitly promote cross-domain generalization (across age groups, camera setups, and environments), so that learned representations remain robust beyond the dataset itself.

In summary, the next generation of gaze datasets should evolve from isolated, 2D-only annotations toward multi-layer, semantically grounded, and actively curated collections.

11.2 Leveraging VLMs for Contextual Gaze Reasoning

Recent advances in large vision–language models (VLMs) open new possibilities for enriching gaze reasoning with contextual knowledge. Whereas current gaze models rely mainly on geometric and saliency cues, VLMs trained on image and text data implicitly encode relationships between people, objects, and actions. Integrating such models could help shift gaze prediction from superficial pattern matching and shortcut behavior toward deeper contextual understanding.

Zero-shot contextual cue extraction. Gupta et al. (2024c) recently demonstrated that VLMs such as BLIP-2 can recognize rich contextual cues such as speaking, manipulating, or interacting, even in a zero-shot manner. When incorporated into a transformer-based gaze model, these cues improved generalization. By querying the model with simple textual prompts (e.g. "Is the person talking?" or "Is the person holding an object?"), VLMs can provide auxiliary features that guide gaze reasoning toward socially and semantically relevant entities rather than geometrically salient ones. This helps mitigate the shortcut learning behavior observed in standard gaze models, where predictions often default to the nearest or largest salient object.

Semi-supervised and pseudo-annotated supervision. Complementary work by Miao et al. (2024) showed that VLMs can also assist in dataset creation. By prompting a pretrained VQA model with gaze-related questions, they obtained Grad-CAM heatmaps that roughly localize gaze targets, which were then refined using a diffusion model trained on a small labeled subset. The resulting pseudo-annotations enabled semi-supervised training that achieved

performance comparable to fully supervised models using only half the labels. This line of work highlights how VLM priors can provide high-quality weak supervision, bridging the gap between large unlabeled datasets and limited human annotations.

Outlook. Looking ahead, VLMs could act as *context providers* within gaze architectures; offering scene-level hypotheses ("who is speaking?", "what action is ongoing?") that complement geometric reasoning. They may also serve as *annotation agents*, generating descriptive scene summaries or question–answer pairs to bootstrap semantic supervision. The main challenges lie in: (i) maintaining spatial grounding since VLMs reason textually but lack pixel-level precision; and (ii) injecting that context in the main gaze architecture, effectively aligning the scene feature space of gaze models with language-driven representations. Addressing these challenges will enable models that not only predict *where* people look but also *why*.

11.3 Temporal Reasoning

In theory, temporal context should greatly aid gaze understanding. Time can provide smoothing, causal disambiguation, and intent inference. For instance, a single frame may make it unclear whether a person looks at a glass or a nearby smartphone; a few frames later, when they reach for the glass, the correct interpretation becomes obvious. Figure 11.1 illustrates this scenario.

Temporal information in practice. Empirically, however, the benefits of explicit temporal modeling remain modest. As shown in the work of Chong et al. (2020b) and our own work Gupta et al. (2024b), complex recurrent or transformer-based temporal modules often yield only marginal gains compared to frame-wise models. In many cases, simple temporal filtering or lightweight smoothing modules applied to the gaze token sequence before decoding might provide comparable consistency with far less complexity.

Why gaze is different. Unlike other computer vision tasks like pose estimation, gaze does not obey strong physical continuity constraints. Body joints have bounded velocity; gaze shifts can occur abruptly—within a fraction of a second—moving from one side of the room to the opposite side. Consequently, over-smoothing can erase meaningful saccades or transient glances. The challenge is therefore not to enforce continuity, but to learn when to smooth and when to allow discontinuity.

Outlook. The limited success of temporal modeling may reflect the insufficiency of current datasets rather than a lack of utility in time itself. In order to truly highlight the importance of temporal disambiguation, we might need to curate a dataset specifically designed to expose situations where the instantaneous gaze target is ambiguous, but becomes evident over time or from past context. With longer, annotated video sequences that capture activities, interactions, and multimodal cues, temporal reasoning could evolve into a more robust and interpretable component of gaze prediction.



Figure 11.1: Illustration of temporal disambiguation in gaze prediction. At time t, the child's gaze target is ambiguous among several possible objects in the scene. At time $t + \Delta t$, when the child reaches for the glass, the intended target becomes evident. Temporal context thus helps resolve gaze uncertainty by linking visual attention with subsequent actions, highlighting the value of temporal reasoning in understanding gaze behavior.

11.4 Native 3D Gaze Following

Despite the inherently three-dimensional nature of visual attention, almost all existing gaze following benchmarks and models remain confined to 2D image-space reasoning. Even methods that incorporate 3D scene information still solve the task in the 2D image space. Adopting an explicit 3D representation of gaze provides a unified formulation of the task, where gaze point estimation and in/out classification become a single problem. It also allows metric evaluation of gaze distance and naturally accommodates occlusions or targets lying outside the visible frame.

From multi-view to metric 3D prediction. Recent efforts like Miao et al. (2025); Hu et al. (2023, 2022a) introduced calibrated multi-camera datasets and depth-annotated indoor scenes to predict a 2D gaze target consistent across views or with the depth map. These works mark an important step from monocular 2D to geometric 3D reasoning. However, their architectures still treat the problem as a 2D localization task with an auxiliary depth or cross-view component. A true 3D formulation would instead predict a volumetric gaze distribution in metric space and project it back to the image when needed.

Metric-scale volumetric representations. An instructive analogy can be drawn from 3D human pose estimation. The MeTRo representation (Sárándi et al., 2020) defines heatmaps directly in metric 3D space rather than image-aligned 2.5D volumes, thereby resolving scale ambiguity and enabling predictions beyond the image boundaries. Adapting this concept to gaze following would allow a model to predict a full 3D gaze heatmap in world coordinates. The in/out classification would then emerge naturally: if the most probable gaze point lies outside the image boundaries after projection, the person is looking out of frame. Such a formulation would not only simplify the learning objective but also make the output physically interpretable and consistent across viewpoints.



Figure 11.2: Example of a reconstructed 3D indoor scene using Gaussian splatting. Such representations capture both geometry and appearance, enabling spatially consistent reasoning about gaze and object relations within the 3D environment.

Integrating 3D geometry and semantics. Beyond geometry, 3D gaze prediction requires a structured scene representation that can be queried by each observer. In constrained environments (e.g. homes, classrooms, or labs), static 3D reconstructions can be built once and updated occasionally to account for moving objects. Recent progress in scene-centric 3D vision-language models, such as GaussianVLM (Halacheva et al., 2025), shows how to embed language-aligned features directly within 3D Gaussian splats. This approach combines geometry, appearance, and semantics into a single queryable representation, enabling embodied agents to reason about the scene in a unified spatial-linguistic space. For gaze following, such representations could help solve the task within a consistent 3D world model (*cf.* Figure 11.2).

Outlook. A native 3D formulation of gaze following promises several benefits: (i) unified treatment of spatial and visibility reasoning; (ii) compatibility with depth-augmented sensors and multi-camera setups; and (iii) seamless integration with embodied, language-aware systems. Achieving this will require advances in 3D scene representation and learning paradigms that tightly couple geometry with semantics, ultimately enabling true 3D gaze target prediction.

11.5 Beyond Gaze: Joint Modeling of Social and Motor Behavior

Gaze understanding has traditionally been approached as a standalone problem, yet in real-world settings, gaze is not an isolated signal but one component of a broader behavioral system that integrates vision, motion, and social intent. To move toward more holistic human-centric models, future research must consider how gaze interacts with other motor and social behaviors that co-occur in natural scenes.

Gaze as part of coordinated behavior. Gaze behavior rarely occurs in isolation. In natural settings, it is tightly coupled with body posture, hand movements, and social cues such as speech and gesture. Studies in psychology and cognitive neuroscience have shown that gaze, gesture, and action are coordinated both temporally and functionally, forming a unified perceptual–motor system that guides attention, supports communication, and reveals intention. Eye movements typically precede manual actions, serving as a planning mechanism for object

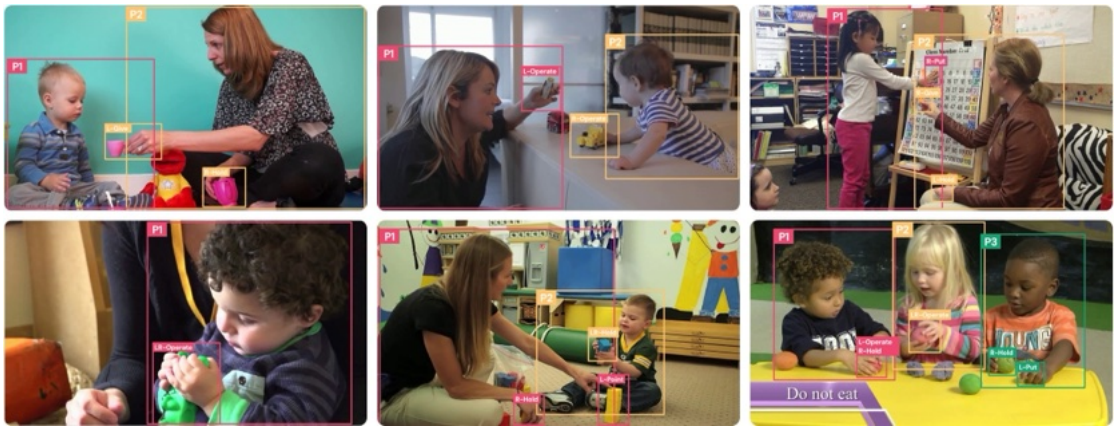


Figure 11.3: Samples from the ChildPlay-Hand dataset. Each frame is annotated with body and object bounding boxes, as well as interaction labels for the left and right hands. The annotations correspond to the same individuals previously labeled for gaze, enabling consistent joint modeling of gaze and hand–object interactions.

manipulation and goal-directed behavior (Land, 2006; Pelz and Canosa, 2001; Huang et al., 2015). Similarly, in social contexts, gaze works in concert with gestures and facial expressions to regulate interaction, signal turn-taking, and facilitate joint attention (Cañigueral and Hamilton, 2019; Hessels et al., 2024). These findings underscore that understanding gaze entails understanding its coordination with other expressive and motor channels.

Coupling gaze with motor and social tasks. Inspired by these behavioral insights, recent work in computer vision has started to integrate gaze with other modalities of human behavior. Huang et al. (2020) proposed the Mutual Context Network, an egocentric model that jointly predicts gaze direction and ongoing actions, showing that visual attention and motor intention can be estimated more accurately when learned together. Hu et al. (2024b) extended this idea to third-person settings, demonstrating that full-body pose dynamics can serve as a strong prior for gaze prediction. Their Pose2Gaze model captures the coordination between eye and body movements during daily activities, highlighting the role of motor context in disambiguating gaze targets. More recently, Özdel et al. (2024) showed that incorporating gaze signals into a graph-based action anticipation network improves the ability to forecast future human actions. Their results empirically confirm what has long been known in behavioral science: where people look reveals what they intend to do next.

Outlook. These advances point toward a broader research direction in which gaze is not treated as an isolated endpoint but as part of a coherent perceptual–motor process. Future models should jointly learn to predict gaze, object interaction, body pose, and social context within a unified architecture. Such a formulation would enable models to capture bidirectional dependencies: actions constrain plausible gaze targets, while gaze reveals intention and facilitates anticipation of future behavior. Our *ChildPlay-Hand* dataset (Farkhondeh et al., 2024) represents an initial step in this direction, coupling gaze following annotations with detailed hand–object interaction labels (*cf.* Figure 11.3). This combination enables studying eye–hand coordination in naturalistic scenes, paving the way for models that reason jointly

about where people look, what they manipulate, and how attention dynamically unfolds during social and motor activities.

11.6 Applications and Societal Impact

The usefulness of gaze following is deeply dependent on the context in which it is applied. As discussed throughout this thesis, gaze prediction models are most informative in proximal, interactional, and socially grounded scenarios, settings where multiple agents share a physical or perceptual space and where understanding attention dynamics is key to interpreting behavior. In such contexts, gaze provides insight into *what* people attend to, *why* they do so, and *how* attention coordinates with social and motor cues. By contrast, distant or surveillance-like settings offer little interpretive value, as the semantic and relational meaning of gaze dissipates when people are far away and the range of possible targets grows large.

Within this framework, several application domains stand to benefit from advances in gaze following. Below, we highlight five representative areas where gaze reasoning can offer tangible impact, both scientifically and societally.

Autism and Neurodevelopmental Assessment. In developmental psychology and clinical practice, gaze following serves as a non-invasive window into social cognition. Impaired gaze behavior, such as reduced joint attention or atypical responses to others' gaze, is one of the earliest observable markers of autism spectrum disorder (ASD). Automatic analysis of gaze and attention patterns during naturalistic interactions can therefore support early screening, progress monitoring, and therapy personalization. Beyond diagnosis, gaze-based metrics can quantify responsiveness and engagement in social therapy sessions, enabling objective longitudinal tracking of improvement. The fine-grained modeling of gaze dynamics proposed in this thesis could help operationalize these behavioral constructs, transforming qualitative observations into measurable indicators for clinical use.

Human–Agent Interaction. Gaze following lies at the heart of intuitive collaboration between humans and intelligent agents. Whether embodied in the physical world (robots) or in virtual environments (avatars, AR assistants), agents benefit from understanding *what* people attend to and *why*. In physical settings, robots use gaze to infer human intent, anticipate motion, and ensure safety in shared workspaces. In virtual and mixed-reality contexts, gaze-based reasoning allows avatars or virtual assistants to establish joint attention, resolve referential ambiguity, and adapt conversational flow. In both cases, the ability to follow human gaze transforms the agent from a reactive system into an attentive partner capable of participating in coordinated, socially aware interaction.

Education and Learning Analytics. In educational settings, gaze is a direct indicator of attention, engagement, and understanding. Predicting where students look and how attention shifts among peers, teachers, and learning materials provides valuable cues for adaptive teaching

and collaborative learning analytics. In small-group or one-on-one tutoring scenarios, multi-person gaze following can help detect who is following the explanation, who is disengaged, and when joint attention is achieved. Such tools could eventually support automatic feedback systems for educators, enabling more personalized and responsive learning environments.

Social and Behavioral Research. Gaze following also offers methodological value in behavioral science and social psychology. Quantifying how attention is distributed and shared within groups provides a powerful means to study leadership, coordination, and communication strategies. Automatic gaze estimation can complement traditional coding of behavior in lab or field studies, enabling fine-grained analyses of turn-taking, attention flow, and interaction structure. This perspective aligns with the broader view of gaze as a social signal that reflects and shapes group dynamics. We successfully applied our gaze following models to help one of our research partners study the role of joint attention in language learning in children (Dickerman et al., 2025). The partner had access to several hours of observational recordings and sought to annotate gaze behavior throughout. Manual coding of such signals would have been prohibitively time-consuming; our models provided an efficient way to scale their analysis, trading off the precision of manual labeling for the breadth and consistency of an automated pipeline.

Marketing and Human Factors in Retail. In applied behavioral research, gaze following can illuminate how people attend to products, interfaces, or other individuals in small-scale, interactive environments. For example, analyzing mutual gaze and shared attention between customers and staff can reveal engagement, trust, and decision-making patterns in retail or service contexts. In product and interface design, understanding how attention flows across interactive elements enables more intuitive layouts and improved user experience. When combined with semantic understanding, gaze target prediction thus provides a practical framework for measuring attention and interaction in real-world human-centered design.

In all these domains, the relevance of gaze following arises from its ability to decode the social and cognitive meaning of attention in context, not from geometric accuracy. As models evolve to reason jointly about gaze and interaction, their applications will extend beyond vision and robotics, toward a broader understanding of human behavior in natural environments.

12 Conclusion

This thesis set out to answer a fundamental question: *How can we design computational models that perceive and reason about human gaze in natural, exo-centric scenes in a way that captures the spatial, social, and semantic principles that underlie human attention?* Across three parts—Foundations, Methods, and Discussion—we explored this question from the perspectives of data, modeling, and interpretation.

The first part established the foundations of gaze understanding. It introduced the conceptual and methodological landscape of gaze following, tracing its origins across psychology, computer vision, and social cognition. It then reviewed the evolution of datasets and benchmarks, highlighting persistent gaps in scale, diversity, and annotation quality that limit generalization. The section on experimental practices consolidated the common metrics, architectural conventions, preprocessing steps, and training routines used in the field to provide a coherent basis for comparison across studies. Finally, an exploratory analysis examined the true content of existing benchmarks, showing that most video-based datasets, despite their size in frames, remain narrow in scene diversity and subsumed within a single large-scale static dataset. These findings underscore the need for larger, richer, more representative, and temporally consistent gaze corpora.

Building on this groundwork, the second part proposed a series of advances that progressively redefined how gaze following can be modeled. Chapter 5 introduced *multimodal and geometrically grounded reasoning*, emphasizing that gaze prediction is inherently a 3D spatial problem requiring consistency between appearance, depth, and field-of-view. Chapter 6 addressed *efficiency and scalability* through the Sharingan architecture, showing that multi-person gaze following can be achieved with a single transformer forward pass, opening the door to real-time and socially contextual inference. This was made possible by the introduction of the *gaze token*, a compact representation of gaze information derived from processing a high-resolution head crop together with its location in the original scene. This concept proved central, as it enabled the model to encode person-specific information efficiently and was later reused across all subsequent architectures. Chapter 7 extended this reasoning toward *social gaze behavior*, unifying tasks such as looking-at-humans, mutual gaze, and shared

attention under that common token-based formulation. Chapter 8 moved beyond localization to *semantic grounding*, where gaze targets are recognized not only by their position but also by their meaning, linking visual attention to semantic understanding. Finally, Chapter 9 explored *predictive uncertainty*, highlighting the probabilistic nature of attention and providing metrics to quantify model confidence and multimodality. Together, these contributions outline a coherent modeling trajectory that progressively integrates geometric reasoning, social context, semantic grounding, and uncertainty modeling, culminating in architectures that combine most of these dimensions within a unified and extensible framework.

The third part turned from technical contributions to broader reflections, examining the limitations, ethical dimensions, and future directions of gaze following research. It discussed the constraints that persist across data, models, and evaluation protocols, as well as the tension between data utility and privacy when working with human-centered visual information. These reflections motivated a shift toward more responsible and context-aware research practices, emphasizing the need for richer and ethically curated datasets, privacy-preserving learning paradigms, and evaluation frameworks that better capture the complexity of human attention. Finally, the discussion looked ahead to emerging directions such as the use of vision–language models for contextual reasoning, the pursuit of native 3D and temporal modeling of gaze, and the integration of gaze with broader social and motor behaviors. Together, these perspectives reframe gaze following not as an isolated perceptual task, but as a stepping stone toward computational models that approach the depth and nuance of human social understanding.

In essence, this thesis repositions gaze following as a bridge between vision and social understanding. By grounding perception in geometry, contextualizing it through social reasoning, and enriching it with semantics and uncertainty, this work advances toward models that go beyond locating gaze targets to interpreting the structure and meaning of visual attention in social contexts. This progression—from pixels to attention, and from attention to meaning—marks a step toward a new generation of socially intelligent vision systems capable of understanding and engaging with humans in natural environments.

Bibliography

al Haytham, I. (1989). *Book of Optics (Kitāb al-Manāir)*. University of Chicago Press. English translation by A.I. Sabra, originally c. 1021.

Archer-Hind, R. D. et al. (1888). *The Timaeus of Plato*. Macmillan.

Ba, S. O. and Odobez, J.-M. (2010). Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116.

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. (2022). Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer.

Bai, C., Kumar, S., Leskovec, J., Metzger, M., Nunamaker Jr, J. F., and Subrahmanian, V. (2019). Predicting the visual focus of attention in multi-person discussion videos. In *IJCAI*, pages 4504–4510.

Bao, J., Liu, B., and Yu, J. (2022). Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135.

Bayliss, A. P., Bartlett, J., Naughtin, C. K., and Kritikos, A. (2011). A direct link between gaze perception and social attention. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3):634.

Belkada, Y., Bertoni, L., Caristan, R., Mordan, T., and Alahi, A. (2021). Do pedestrians pay attention? eye contact detection in the wild. *arXiv preprint arXiv:2112.04212*.

Bishop, C. M. (1994). Mixture density networks. Technical report, Aston University.

Boluk, N. and Kose, H. (2024). Evaluating gaze detection for children with autism using the childplay-r dataset. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5. IEEE.

Brody, S., Alon, U., and Yahav, E. (2021). How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*.

Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental science*, 8(6):535–543.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2018). What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757.

Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., and Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, 40(8):1129–1138.

Cañigueral, R. and Hamilton, A. F. d. C. (2019). The role of eye gaze during natural social interactions in typical and autistic people. *Frontiers in psychology*, 10:560.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Chang, F., Zeng, J., Liu, Q., and Shan, S. (2023). Gaze pattern recognition in dyadic communication. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–7.

Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., and Deng, J. (2018). Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE.

Chen, J., Konrad, J., and Ishwar, P. (2018). Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1570–1579.

Chen, J., Yang, Z., and Zhang, L. (2023). Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything.

Chen, Q. and Hong, Y. (2024). Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian conference on computer vision*, pages 2404–2420.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., and Qiao, Y. (2022). Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299.

Choi, J., Chun, D., Kim, H., and Lee, H.-J. (2019). Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 502–511.

Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R. M., Rozga, A., and Rehg, J. M. (2017). Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20.

Chong, E., Clark-Whitney, E., Southerland, A., Stubbs, E., Miller, C., Ajodan, E. L., Silverman, M. R., Lord, C., Rozga, A., Jones, R. M., et al. (2020a). Detection of eye contact with deep neural networks is as accurate as human experts. *Nature communications*, 11(1):6386.

Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., and Rehg, J. M. (2018). Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398.

Chong, E., Wang, Y., Ruiz, N., and Rehg, J. M. (2020b). Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5396–5406.

de Belen, R. A., Mohammadi, G., and Sowmya, A. (2023). Temporal understanding of gaze communication with gazetransformer. In *NeuRIPS 2023 Workshop on Gaze Meets ML*.

De Belen, R. A. J., Bednarz, T., Sowmya, A., and Del Favero, D. (2020). Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry*, 10(1):333.

de Belen, R. A. J., Mohammadi, G., and Sowmya, A. (2025). Gazedetr: Gaze detection using disentangled head and gaze representations. *arXiv preprint arXiv:2508.12966*.

Degutyte, Z. and Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings. *Frontiers in Psychology*, 12:616471.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dickerman, M., Gupta, A., Tafasca, S., Zhang, X., Odobez, J.-M., and Stoll, S. (2025). Automatic detection of the visual gaze components of joint attention in naturalistic, observational data. In *Proceedings of the 49th Annual Boston University Conference on Language Development*. Cascadilla Press.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Doosti, B., Chen, C.-H., Vemulapalli, R., Jia, X., Zhu, Y., and Green, B. (2021). Boosting image-based mutual gaze detection using pseudo 3d gaze. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1273–1281.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dwivedi, V. P. and Bresson, X. (2020). A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Fan, L., Chen, Y., Wei, P., Wang, W., and Zhu, S.-C. (2018). Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468.

Fan, L., Wang, W., Huang, S., Tang, X., and Zhu, S.-C. (2019). Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733.

Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., and Zhai, G. (2021). Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11390–11399.

Farkhondeh, A., Tafasca, S., and Odobez, J.-M. (2024). Childplay-hand: A dataset of hand manipulations in the wild. In *European Conference on Computer Vision*, pages 101–118. Springer.

Farroni, T., Csibra, G., Simion, F., and Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National academy of sciences*, 99(14):9602–9605.

Franchak, J. M., Heeger, D. J., Hasson, U., and Adolph, K. E. (2016). Free viewing gaze behavior in infants and adults. *Infancy*, 21(3):262–287.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gao, C., Zou, Y., and Huang, J.-B. (2018). ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.

Gkioxari, G., Girshick, R., Dollár, P., and He, K. (2018). Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367.

Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Guan, J., Yin, L., Sun, J., Qi, S., Wang, X., and Liao, Q. (2019). Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*, pages 502–513. Springer.

Guo, H., Hu, Z., and Liu, J. (2022). Mgtr: End-to-end mutual gaze detection with transformer. In *Proceedings of the Asian Conference on Computer Vision*, pages 1590–1605.

Gupta, A., Dollar, P., and Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

Gupta, A., Tafasca, S., Chutisilp, N., and Odobez, J.-M. (2024a). A unified model for gaze following and social gaze prediction. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE.

Gupta, A., Tafasca, S., Farkhondeh, A., Vuillecard, P., and Odobez, J.-M. (2024b). Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. *Advances in Neural Information Processing Systems*, 37:15646–15673.

Gupta, A., Tafasca, S., and Odobez, J.-M. (2022). A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5041–5050.

Gupta, A., Vuillecard, P., Farkhondeh, A., and Odobez, J.-M. (2024c). Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 615–624.

Gupta, S. and Malik, J. (2015). Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.

Halacheva, A.-M., Zaech, J.-N., Wang, X., Paudel, D. P., and Van Gool, L. (2025). Gaussian-vlm: Scene-centric 3d vision-language models using language-aligned gaussian splats for embodied reasoning and beyond. *arXiv preprint arXiv:2507.00886*.

Hashemi, J., Dawson, G., Carpenter, K. L., Campbell, K., Qiu, Q., Espinosa, S., Marsan, S., Baker, J. P., Egger, H. L., and Sapiro, G. (2018). Computer vision analysis for quantification of autism risk behaviors. *IEEE Transactions on Affective Computing*, 12(1):215–226.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hempel, T., Jung, M., Abdelrahman, A. A., and Al-Hamadi, A. (2024). Nitec: Versatile hand-annotated eye contact dataset for ego-vision interaction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4437–4446.

Hermann, K. L., Mobahi, H., Fel, T., and Mozer, M. C. (2023). On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*.

Hessels, R. S., Li, P., Balali, S., Teunisse, M. K., Poppe, R., Niehorster, D. C., Nyström, M., Benjamins, J. S., Senju, A., Salah, A. A., et al. (2024). Gaze-action coupling, gaze-gesture coupling, and exogenous attraction of gaze in dyadic interactions. *Attention, Perception, & Psychophysics*, 86(8):2761–2777.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Hogan, E. A. (2001). The attention economy: understanding the new currency of business.

Hou, Y., Zhang, Z., Horanyi, N., Moon, J., Cheng, Y., and Chang, H. J. (2024). Multi-modal gaze following in conversational scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1186–1195.

Hu, S., Sun, H., Wei, D., Sun, X., and Wang, J. (2024a). Continuous heatmap regression for pose estimation via implicit neural representation. *Advances in Neural Information Processing Systems*, 37:102036–102055.

Hu, Z., Xu, J., Schmitt, S., and Bulling, A. (2024b). Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses. *IEEE Transactions on Visualization and Computer Graphics*.

Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S., and Liu, J. (2022a). We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14.

Hu, Z., Yang, Y., Zhai, X., Yang, D., Zhou, B., and Liu, J. (2023). Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8907–8916.

Hu, Z., Zhao, K., Zhou, B., Guo, H., Wu, S., Yang, Y., and Liu, J. (2022b). Gaze target estimation inspired by interactive attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8524–8536.

Huang, C.-M., Andrist, S., Sauppé, A., and Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6:1049.

Huang, Y., Cai, M., Li, Z., Lu, F., and Sato, Y. (2020). Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., and Brox, T. (2018). Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 652–667.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jia, S. and Bruce, N. D. (2020). Revisiting saliency metrics: Farthest-neighbor area under curve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2667–2676.

Jin, T., Lin, Z., Zhu, S., Wang, W., and Hu, S. (2021). Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE.

Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y., and Song, W. (2022). Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924.

Jin, Y., Zhang, L., Yan, S., Fan, B., and Wang, B. (2024). Boosting gaze object prediction via pixel-level supervision from vision foundation model. In *European Conference on Computer Vision*, pages 369–386. Springer.

Joshi, C. K. (2025). Transformers are graph neural networks. *arXiv preprint arXiv:2506.22084*.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921.

Kendall, A., Badrinarayanan, V., and Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.

Kendall, A. and Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5974–5983.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., and Ronneberger, O. (2018). A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31.

Kotseruba, I., Rasouli, A., and Tsotsos, J. K. (2016). Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Labelbox (2025). Labelbox. https://labelbox.com.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lan, X., Hu, Q., Chen, Q., Xue, J., and Cheng, J. (2021a). Hih: Towards more accurate face alignment via heatmap in heatmap. *arXiv preprint arXiv:2104.03100*.

Lan, X., Hu, Q., and Cheng, J. (2021b). Revisting quantization error in face alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1521–1530.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, 25(3):296–324.

Le, Q. V., Smola, A. J., and Canu, S. (2005). Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496.

Li, B., Wang, Q., Barney, E., Hart, L., Wall, C., Chawarska, K., de Urabain, I. S., Smith, T. J., and Shic, F. (2016). Modified dbscan algorithm on oculomotor fixation identification. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 337–338.

Li, C. and Lee, G. H. (2019). Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895.

Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W., and Okumura, M. (2024). A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5879–5899.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022a). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. (2021a). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. (2022b). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Li, Y., Shen, W., Gao, Z., Zhu, Y., Zhai, G., and Guo, G. (2021b). Looking here or there? gaze following in 360-degree images. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3742–3751.

Lian, D., Yu, Z., and Gao, S. (2018). Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer.

Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., and Feng, J. (2020). Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017b). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lin, Z.-Y., Chew, J. Y., van Gemert, J., and Zhang, X. (2025). Gazehta: End-to-end gaze target detection with head-target association. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9447–9454. IEEE.

Lindberg, D. C. (1976). *Theories of Vision from al-Kindi to Kepler*. University of Chicago Press.

Liu, F., Li, K., Zhong, Z., Jia, W., Hu, B., Yang, X., Wang, M., and Guo, D. (2024). Depth matters: Spatial proximity-based gaze cone generation for gaze following in wild. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11):1–24.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. (2023). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luvizon, D. C., Tabia, H., and Picard, D. (2019). Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22.

Makansi, O., Ilg, E., Cicek, O., and Brox, T. (2019). Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7153.

Malik, M. and Isik, L. (2023). Relational visual representations underlie human social interaction recognition. *Nature Communications*, 14(1):7317.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Marin-Jimenez, M. J., Kalogeiton, V., Medina-Suarez, P., and Zisserman, A. (2019). Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485.

Marin-Jimenez, M. J., Kalogeiton, V., Medina-Suarez, P., and Zisserman, A. (2022). Laeo-Net++: Revisiting People Looking at Each Other in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):3069–3081.

Marin-Jimenez, M. J., Zisserman, A., Eichner, M., and Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296.

Mathew, A. M., Khan, A. A., Khalid, T., Al-Tam, F., and Souissi, R. (2025). Leveraging multimodal saliency and fusion for gaze target detection. *arXiv preprint arXiv:2504.19271*.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Miao, Q., Golani, V. R., Xu, J., Dutta, P. P., Hoai, M., and Samaras, D. (2025). Multi-view gaze target estimation. *arXiv preprint arXiv:2508.05857*.

Miao, Q., Graikos, A., Zhang, J., Mondal, S., Hoai, M., and Samaras, D. (2024). Diffusion-refined vqa annotations for semi-supervised gaze following. In *European Conference on Computer Vision*, pages 439–457. Springer.

Miao, Q., Hoai, M., and Samaras, D. (2023). Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 880–889.

Mohammed, O. K., Aggarwal, K., Liu, Q., Singhal, S., Bjorck, J., and Som, S. (2023). Bootstrapping a high quality multilingual multimodal dataset for bletchley. In *Asian Conference on Machine Learning*, pages 738–753. PMLR.

Mundy, P. and Newell, L. (2007). Attention, joint attention, and social cognition. *Current directions in psychological science*, 16(5):269–274.

Munea, T. L., Jembre, Y. Z., Weldegebriel, H. T., Chen, L., Huang, C., and Yang, C. (2020). The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *Ieee Access*, 8:133330–133348.

Nakatani, C., Kawashima, H., and Ukita, N. (2023). Interaction-aware joint attention estimation using people attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10224–10233.

Neumann, D., Spezio, M. L., Piven, J., and Adolphs, R. (2006). Looking you in the mouth: abnormal gaze in autism resulting from impaired top-down modulation of visual attention. *Social cognitive and affective neuroscience*, 1(3):194–202.

Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer.

Nibali, A., He, Z., Morgan, S., and Prendergast, L. (2018). Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*.

Onkhar, V., Bazilinskyy, P., Stapel, J. C., Dodou, D., Gavrila, D., and de Winter, J. C. (2021). Towards the detection of driver–pedestrian eye contact. *Pervasive and Mobile Computing*, 76:101455.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Otsuka, K., Kasuga, K., and Köhler, M. (2018). Estimating visual focus of attention in multi-party meetings using deep convolutional neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 191–199.

Otsuka, K., Yamato, J., Takemae, Y., and Murase, H. (2006). Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In *2006 IEEE International Conference on Multimedia and Expo*, pages 949–952. IEEE.

Özdel, S., Rong, Y., Albaba, B. M., Kuo, Y.-L., Wang, X., and Kasneci, E. (2024). Gaze-guided graph neural network for action anticipation conditioned on intention. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pages 1–9.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Patakin, N., Vorontsova, A., Artemyev, M., and Konushin, A. (2022). Single-stage 3d geometry-preserving depth estimation model training on dataset mixtures with uncalibrated stereo data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1705–1714.

Pelz, J. B. and Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision research*, 41(25-26):3587–3596.

Penchala, S., Kontham, S. R., Bhattacharjee, P., Karami, S., Ghahremani, M., Golilarz, N. A., and Rahimi, S. (2025). Learning in focus: Detecting behavioral and collaborative engagement using vision transformers. *arXiv preprint arXiv:2508.15782*.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188.

Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. (2015). Where are they looking? *Advances in neural information processing systems*, 28.

Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C., et al. (2013). Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3414–3421.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L. (2024). Grounded sam: Assembling open-world models for diverse visual tasks.

Rommel, C., Letzelter, V., Samet, N., Marlet, R., Cord, M., Pérez, P., and Valle, E. (2023). Manipose: Manifold-constrained multi-hypothesis 3d human pose estimation. *arXiv preprint arXiv:2312.06386*.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. (2017). Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., and Rehg, J. M. (2025). Gaze-lle: Gaze target estimation via large-scale learned encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28874–28884.

Sárándi, I., Linder, T., Arras, K. O., and Leibe, B. (2020). Metrabs: metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30.

Scaife, M. and Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*, 253(5489):265–266.

Sciortino, G., Farinella, G. M., Battiato, S., Leo, M., and Distante, C. (2017). On the estimation of children's poses. In *International conference on image analysis and processing*, pages 410–421. Springer.

Seitzer, M., Tavakoli, A., Antic, D., and Martius, G. (2022). On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*.

Shepherd, S. V. (2010). Following gaze: gaze-following behavior as a window into social cognition. *Frontiers in integrative neuroscience*, 4:5.

Shic, F., Bradshaw, J., Klin, A., Scassellati, B., and Chawarska, K. (2011). Limited activity monitoring in toddlers with autism spectrum disorder. *Brain research*, 1380:246–254.

Shields, C. J. et al. (2016). *Aristotle: De Anima*. Oxford University Press.

Siegfried, R. and Odobez, J.-M. (2021). Visual focus of attention estimation in 3d scene with an arbitrary number of targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3153–3161.

Smith, F. B., Kossen, J., Trollope, E., van der Wilk, M., Foster, A., and Rainforth, T. (2024). Rethinking aleatoric and epistemic uncertainty. *arXiv preprint arXiv:2412.20892*.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Song, Y., Wang, X., Yao, J., Liu, W., Zhang, J., and Xu, X. (2024). Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2(1):31.

Stephenson, L. J., Edwards, S. G., and Bayliss, A. P. (2021). From gaze perception to social cognition: The shared-attention system. *Perspectives on Psychological Science*, 16(3):553–576.

Stiefelhagen, R., Yang, J., and Waibel, A. (2001). Estimating focus of attention based on gaze and sound. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–9.

Stirn, A., Wessels, H., Schertzer, M., Pereira, L., Sanjana, N., and Knowles, D. (2023). Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 5593–5613. PMLR.

Sumer, O., Gerjets, P., Trautwein, U., and Kasneci, E. (2020). Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3327–3336.

Syeda, U. H., Zafar, Z., Islam, Z. Z., Tazwar, S. M., Rasna, M. J., Kise, K., and Ahad, M. A. R. (2017). Visual face scanning and emotion perception analysis between autistic and typically developing children. In *Proceedings of the 2017 acm international joint conference on pervasive and ubiquitous computing and proceedings of the 2017 acm international symposium on wearable computers*, pages 844–853.

Tafasca, S., Gupta, A., Bros, V., and Odobez, J.-M. (2024a). Toward semantic gaze target detection. *Advances in neural information processing systems*, 37:121422–121448.

Tafasca, S., Gupta, A., Kojovic, N., Gelsomini, M., Maillart, T., Papandrea, M., Schaer, M., and Odobez, J.-M. (2023a). The ai4autism project: A multimodal and interdisciplinary approach to autism diagnosis and stratification. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 414–425.

Tafasca, S., Gupta, A., and Odobez, J.-M. (2023b). Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946.

Tafasca, S., Gupta, A., and Odobez, J.-M. (2024b). Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2017.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Thompson, S., Gupta, A., Gupta, A. W., Chen, A., and Vázquez, M. (2021). Conversational group detection with graph neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 248–252.

Toaiari, A., Murino, V., Cristani, M., and Beyan, C. (2024). Upper-body pose-based gaze estimation for privacy-preserving 3d gaze target detection. In *European Conference on Computer Vision*, pages 359–376. Springer.

Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., and Guinto, R. (2021). Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133.

Tonini, F., Beyan, C., and Ricci, E. (2022). Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 420–431.

Tonini, F., Dall'Asen, N., Beyan, C., and Ricci, E. (2023). Object-aware gaze target detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21860–21869.

Tonini, F., Dall'Asen, N., Vaquero, L., Beyan, C., and Ricci, E. (2024). Al-gtd: Deep active learning for gaze target detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2360–2369.

Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., and Shen, W. (2022). End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2200. IEEE.

Tu, D., Shen, W., Sun, W., Min, X., Zhai, G., and Chen, C. (2023). Un-gaze: A unified transformer for joint gaze-location and gaze-object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3271–3285.

Varamesh, A. and Tuytelaars, T. (2020). Mixture dense regression for object detection and human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13086–13095.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vondrick, C., Patterson, D., and Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling. *International journal of computer vision*, 101(1):184–204.

Wang, B., Guo, C., Jin, Y., Xia, H., and Liu, N. (2024). Transgop: transformer-based gaze object prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10180–10188.

Wang, B., Hu, T., Li, B., Chen, X., and Zhang, Z. (2022). Gatector: A unified framework for gaze object prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19588–19597.

Wang, S., Yap, K.-H., Ding, H., Wu, J., Yuan, J., and Tan, Y.-P. (2021). Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13475–13484.

Wang, X., Zhang, H., Wang, Z., Nie, W., Yang, Z., Ren, W., Xu, Q., Xu, X., and Liu, H. (2023). Dual regression-enhanced gaze target detection in the wild. *IEEE Transactions on Cybernetics*, 54(1):219–229.

Wei, P., Liu, Y., Shu, T., Zheng, N., and Zhu, S.-C. (2018). Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6801–6809.

Wu, T. and Sugano, Y. (2022). Learning video-independent eye contact segmentation from in-the-wild videos. In *Proceedings of the Asian Conference on Computer Vision*, pages 703–720.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

Xie, Z., Liu, Z., Peng, Z., Wu, W., and Zhou, B. (2025). Vid2sim: Realistic and interactive simulation from video for urban navigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1581–1591.

Yang, J., Li, B., Yang, F., Zeng, A., Zhang, L., and Zhang, R. (2023). Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*.

Yang, Y. and Lu, F. (2024). Gaze target detection based on head-local-global coordination. In *European Conference on Computer Vision*, pages 305–322. Springer.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., and Shen, C. (2021). Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 204–213.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019a). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Yun, S., Jeong, M., Kim, R., Kang, J., and Kim, H. J. (2019b). Graph transformer networks. *Advances in neural information processing systems*, 32.

Zhang, F., Zhu, X., Dai, H., Ye, M., and Zhu, C. (2020). Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102.

Zhang, G., Luo, Z., Yu, Y., Cui, K., and Lu, S. (2022). Accelerating detr convergence via semantic-aligned matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 949–958.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, L., Morgan, M., Bhattacharya, I., Foley, M., Braasch, J., Riedl, C., Foucault Welles, B., and Radke, R. J. (2019). Improved visual focus of attention estimation and prosodic features for analyzing group interactions. In *2019 International Conference on Multimodal Interaction*, pages 385–394.

Zhao, H., Lu, M., Yao, A., Chen, Y., and Zhang, L. (2020). Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100.

Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., and Oliva, A. (2016). Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641.

Zhuang, B., Wu, Q., Shen, C., Reid, I., and van den Hengel, A. (2018). Hcvrd: A benchmark for large-scale human-centered visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.